NATURAL DISASTERS ANALYSIS

Submitted to : Dr. Santosh Singh Rathore

DONE BY :

BHAVANA MEKALA 2021BCS-041

NELLURI PAVITHRA SAI LAKSHMI 2021BCS-049

TABLE OF CONTENTS

INTRODUCTION

- 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.
- Earthquakes, Hurricanes (Cyclones), Tornadoes, Floods, Volcanic eruptions, Tsunamis, Droughts, Blizzards, Landslides, Heatwaves, Ice storms, Hailstorms are some common types of natural disasters.
- This analysis focuses on the period from 1900 to 2021, examining the patterns and consequences of these events.
- The EM-DAT database, a crucial resource, captures a century of disasters with meticulous detail, offering insights into their types, locations, and socio-economic impacts.

DATA COLLECTION

AND

UNDERSTANDING DATASET

DATA COLLECTION

- We have collected our dataset from kaggle which is taken from 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.
- Number of Samples : 16124
- Number of Features : 31

DATASET(BEFORE DATA CLEANING)

	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Disaster	Disaster	Event Na	Country	ISO	Continen	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 Type 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

DATA PREPROCESSING

DATA PRE-PROCESSING

- 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 new attribute End Date.

Handling Missing Values :

- For numerical values replaced them with mean and For categorical values replaced them with a common value or mode.
- Dropped redundant or uninformative columns like Dis No etc.

DATASET AFTER CLEANING

```
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

```python
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)
```

DATASET AFTER CLEANING

```
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%
```

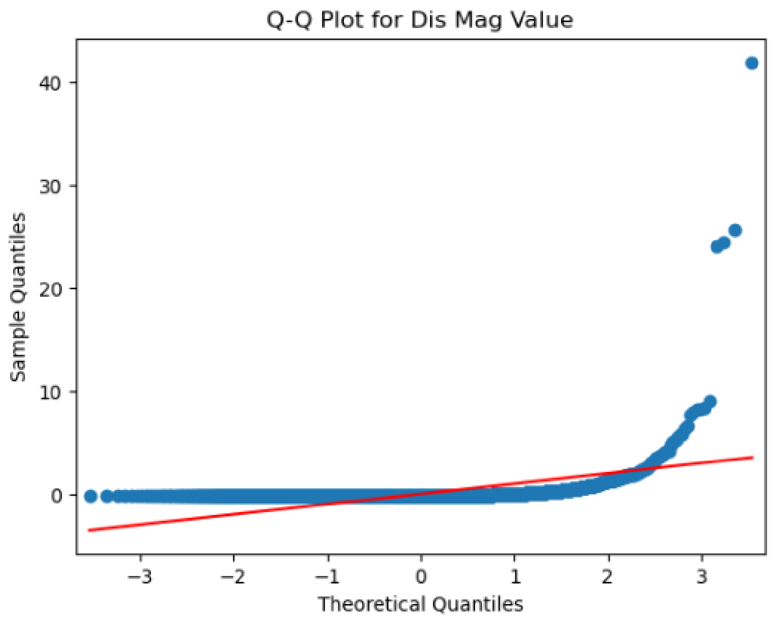
We can observe that before the dataset had 31 columns and now reduced to 21 columns after performing data pre-processing.

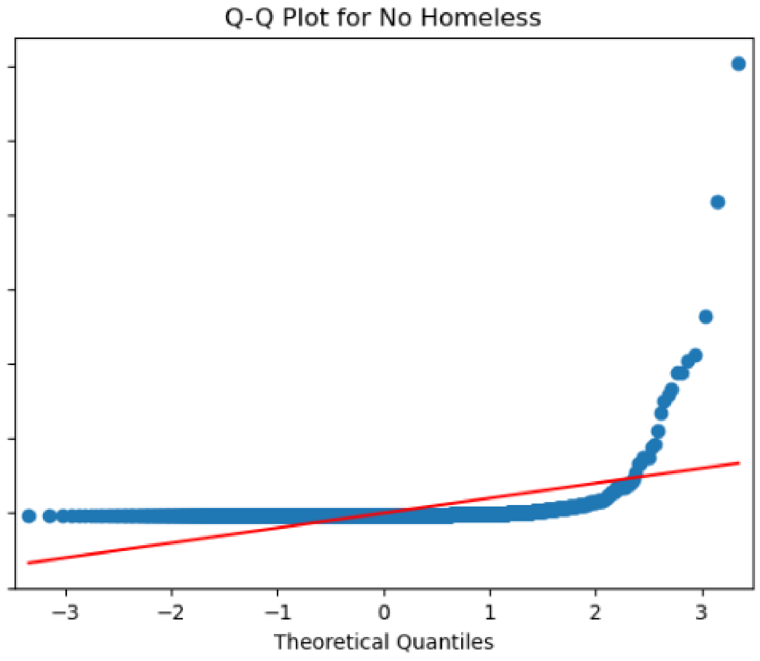
And number of samples reduced from 16124 to 14332.

Attributes

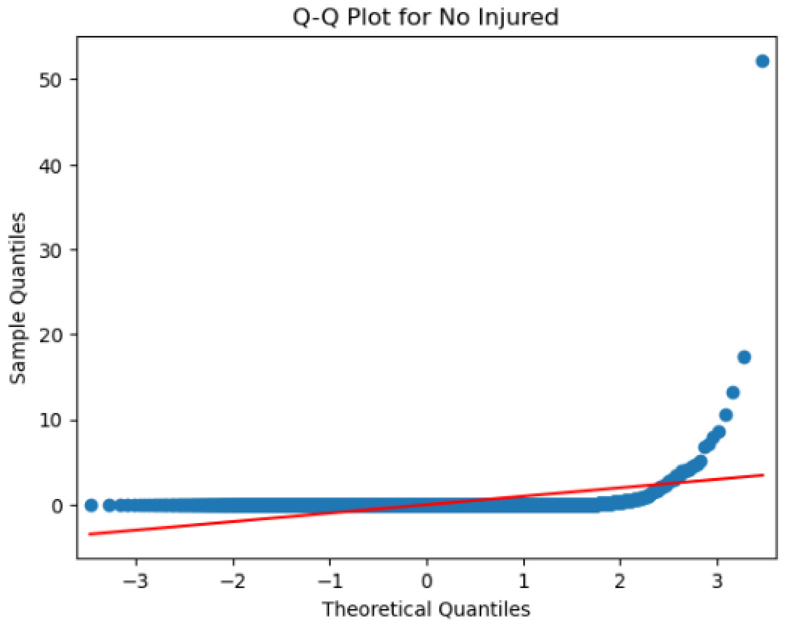
	Name	dtypes
0	Year	int64
1	Disaster Subgroup	object
2	Disaster Type	object
3	Event Name	object
4	Country	object
5	ISO	object
6	Continent	object
7	Location	object
8	Origin	object
9	Associated Dis	object
10	Declaration	object
11	Dis Mag Value	float64
12	Dis Mag Scale	object
13	Latitude	object
14	Longitude	object
15	Local Time	object
16	River Basin	object
17	Start Date	object
18	End Date	object
19	Total Deaths	float64
20	No Injured	float64
21	No Affected	float64
22	No Homeless	float64
23	Total Affected	float64
24	Insured Damages ('000 US$)	float64
25	Total Damages ('000 US$)	float64
26	CPI	float64

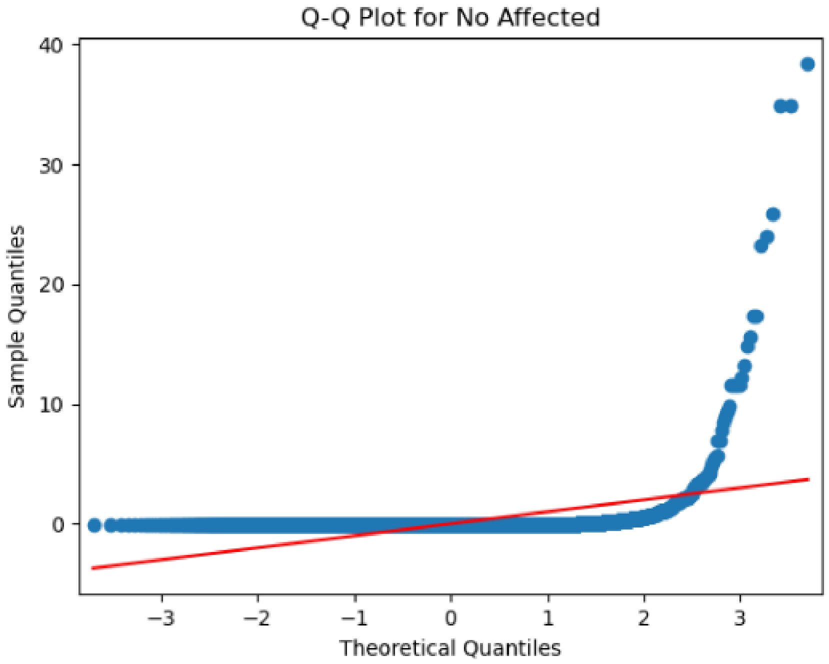
Q-Q PLOTS:


Q-Q Plot for Dis Mag Value
Sample Quantiles
Theoretical Quantiles
40
30
20
10
0
−3
−2
−1
0
1
2
3

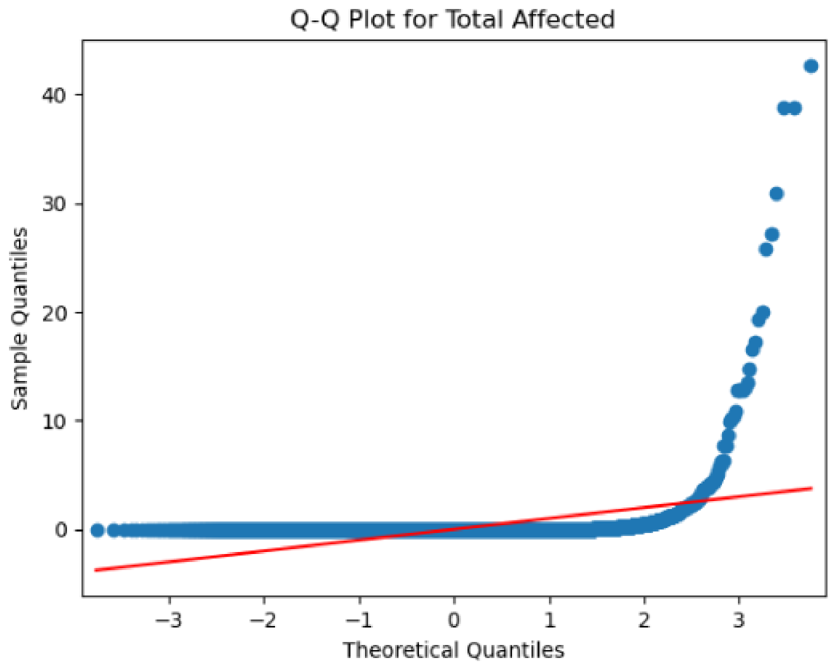

Q-Q Plot for No Homeless
Theoretical Quantiles
−3
−2
−1
0
1
2
3

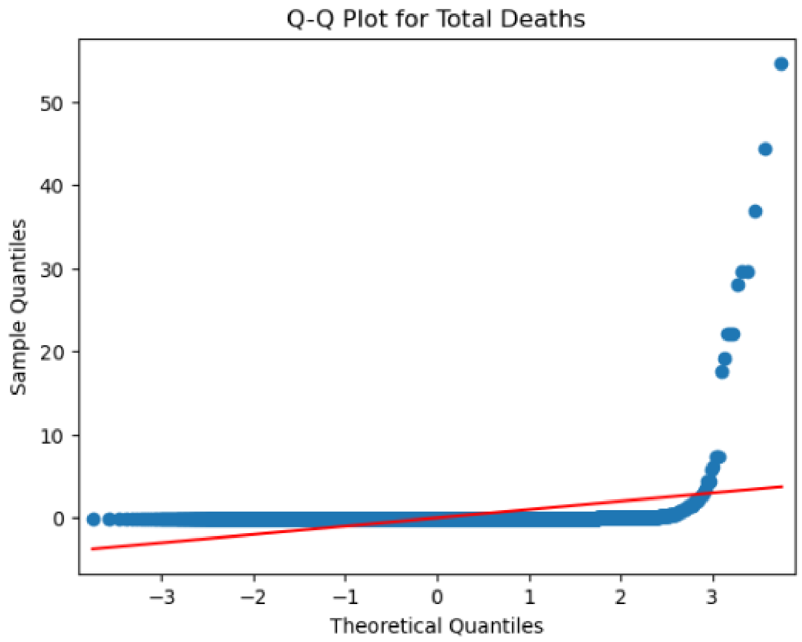
Q-Q PLOTS:


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

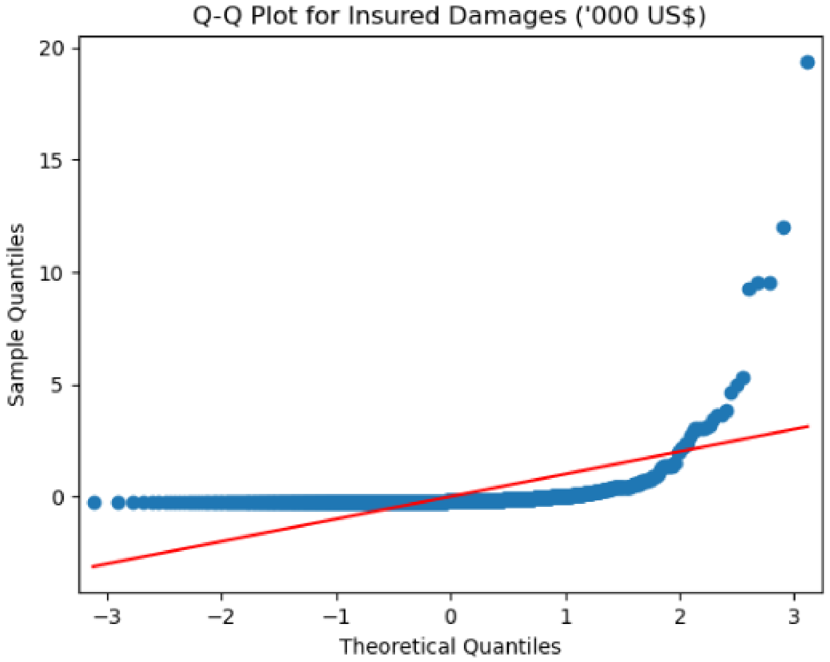

Q-Q Plot for No Affected
Sample Quantiles
Theoretical Quantiles
40
30
20
10
0
−3
−2
−1
0
1
2
3

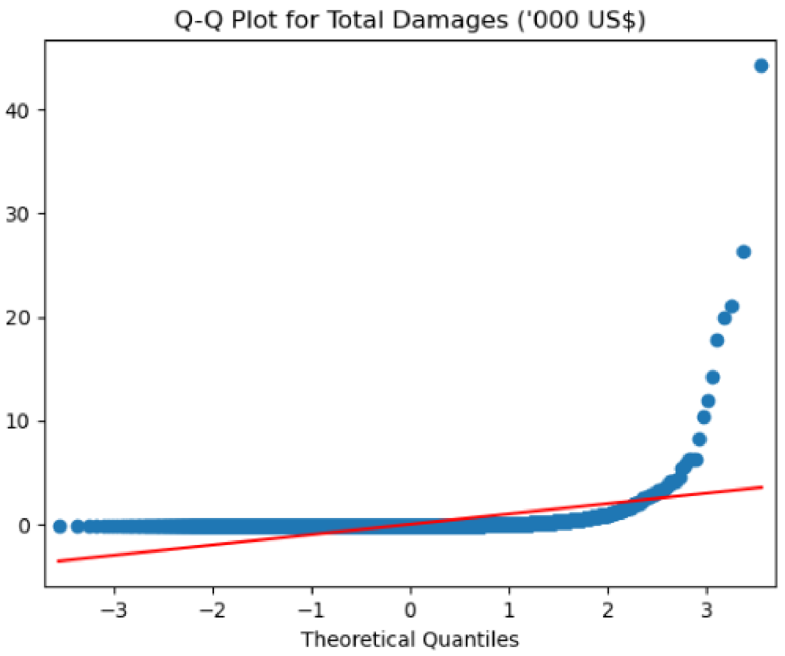
Q-Q PLOTS:


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

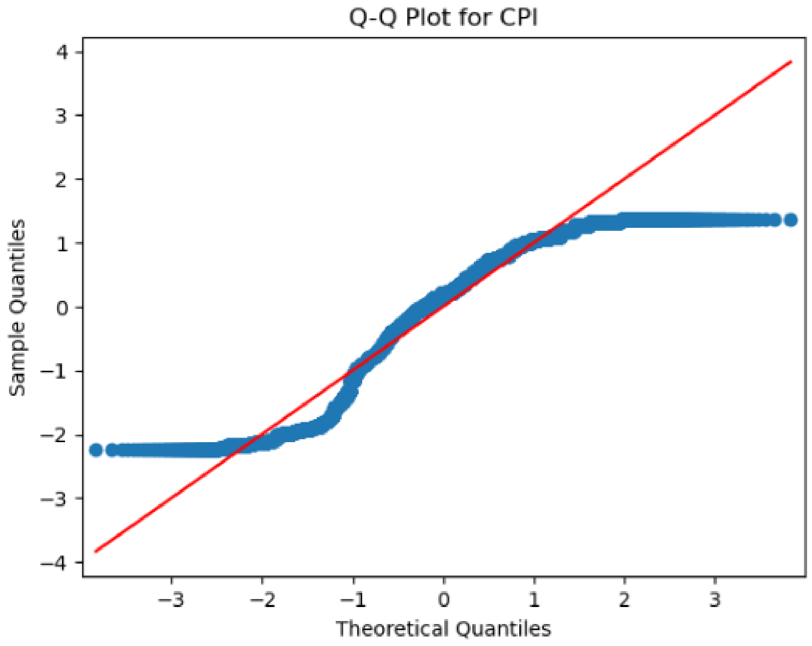

Q-Q Plot for Total Deaths
Sample Quantiles
0
10
20
30
40
50
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Q-Q PLOTS:


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


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

Q-Q PLOTS:


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

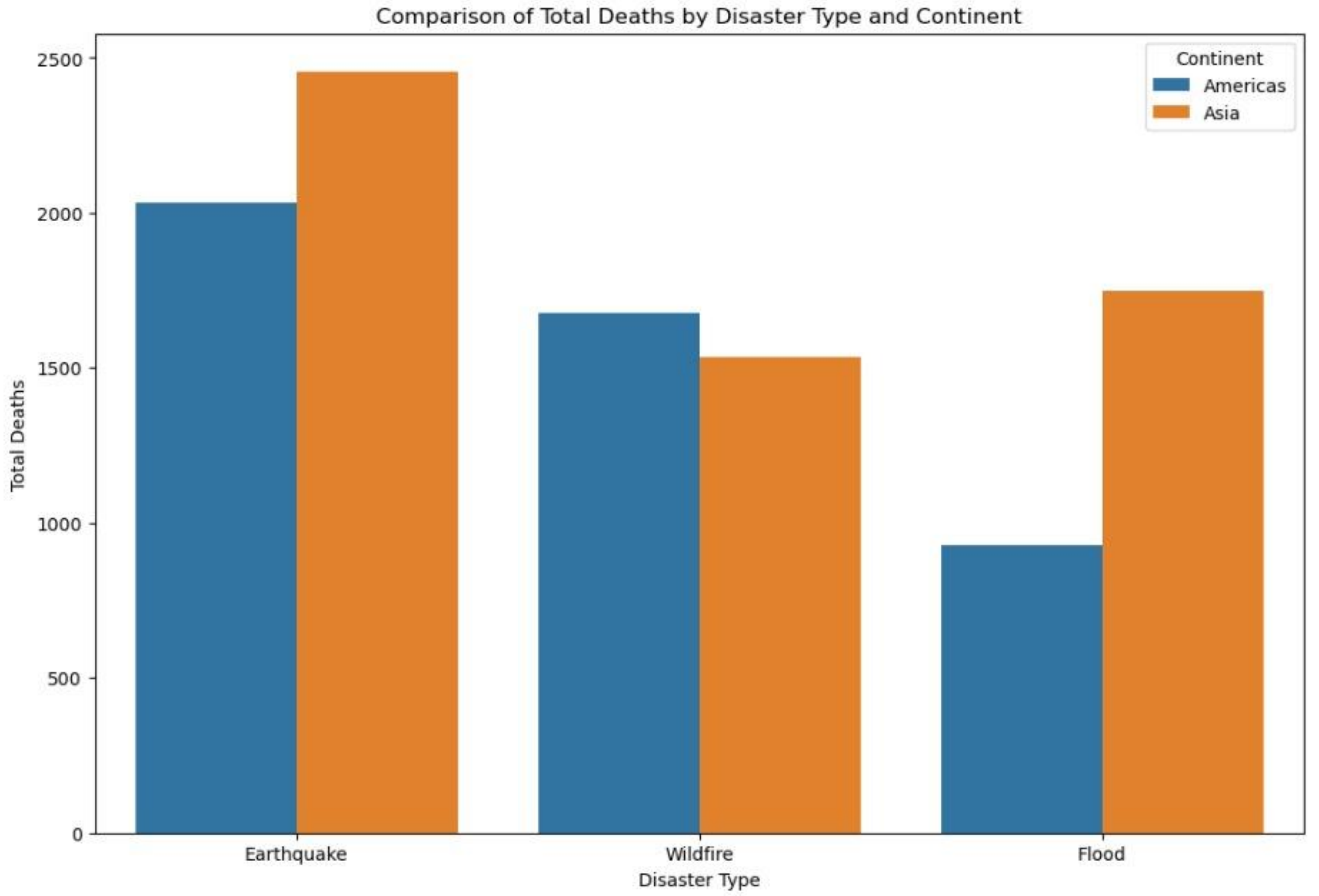
FINDINGS

- We have drawn Q-Q plots to find what tests should be used.
- The Q-Q plot's visual examination reveals that the observed points differ from the predicted straight line, proving that the distribution is not normal.
- A normal distribution may be assumed in statistical analysis, which may be affected by this deviation from normality.
- 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.

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.

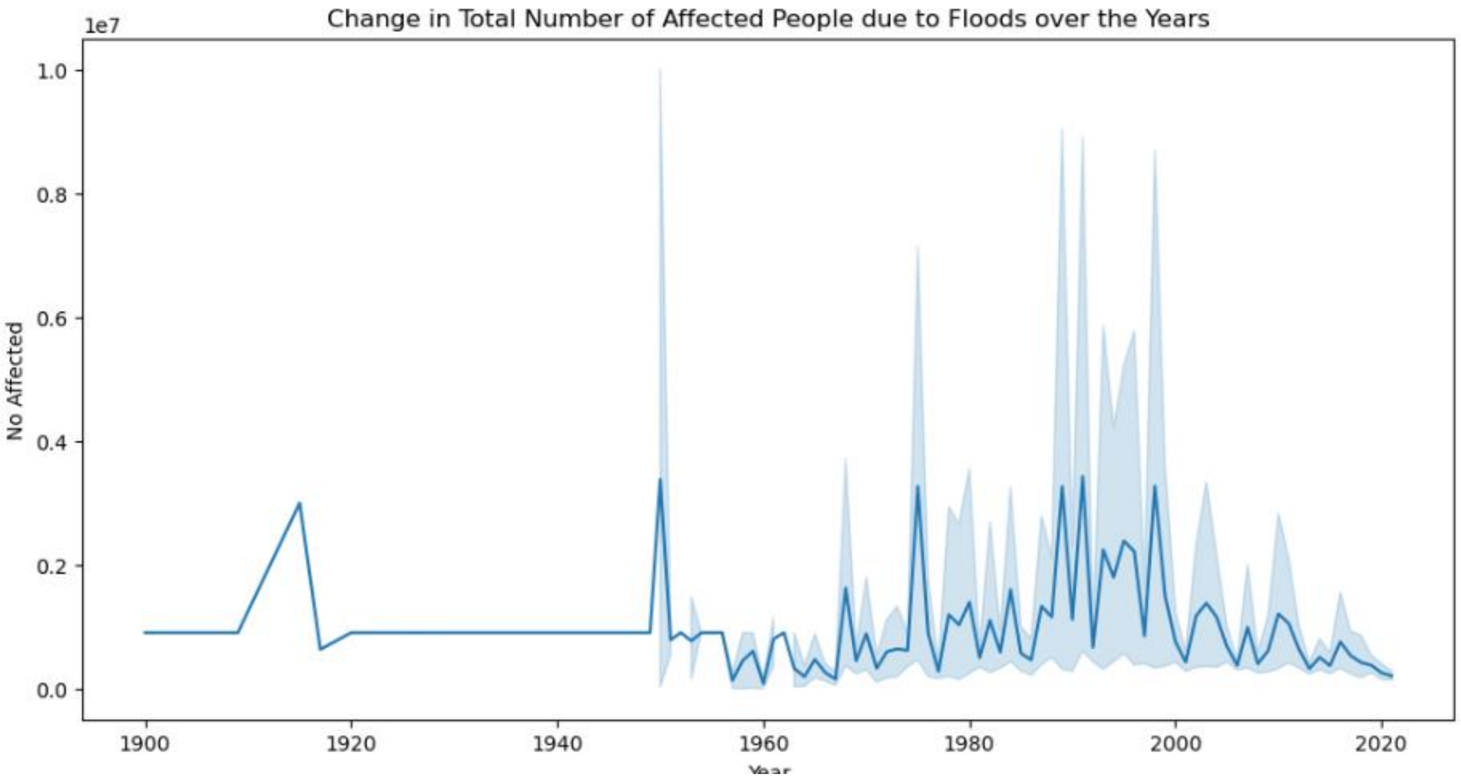
Null Hypothesis (H0): The distributions of total deaths for each disaster subtype are the same across Asia and North America.

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

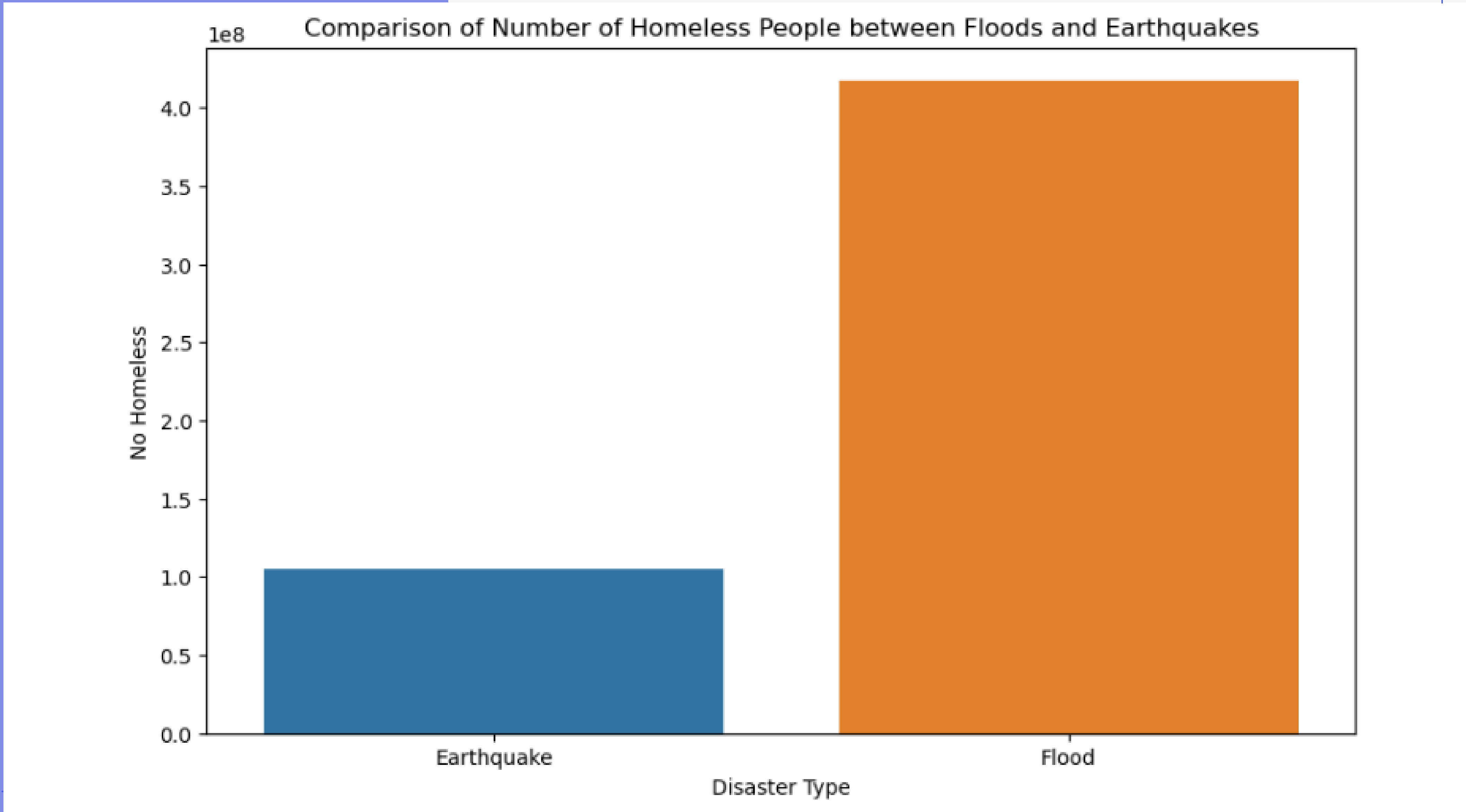
In [18]:
```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import kruskal

# 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'].drc
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.
```

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



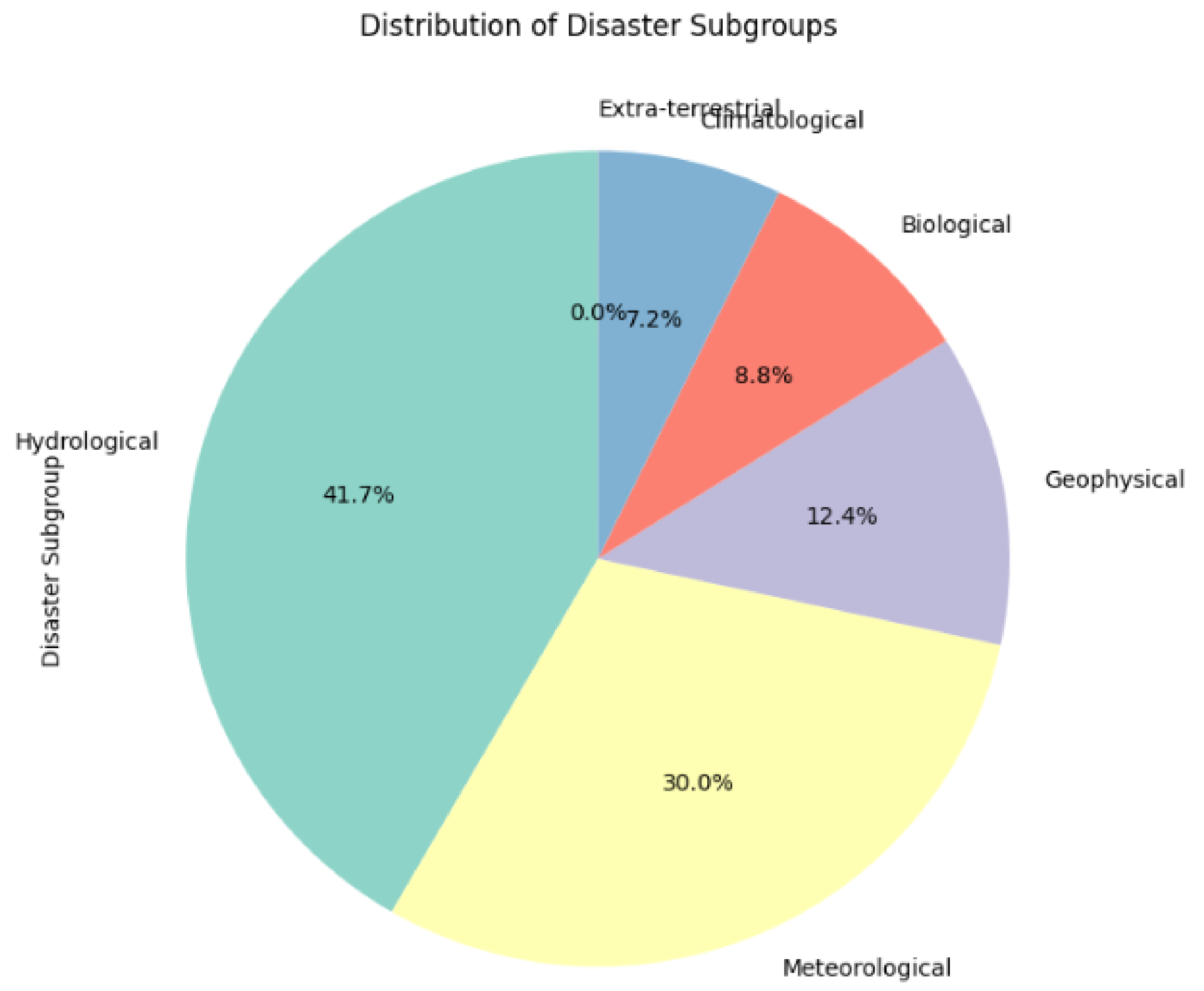
```
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import wilcoxon

# 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()
```



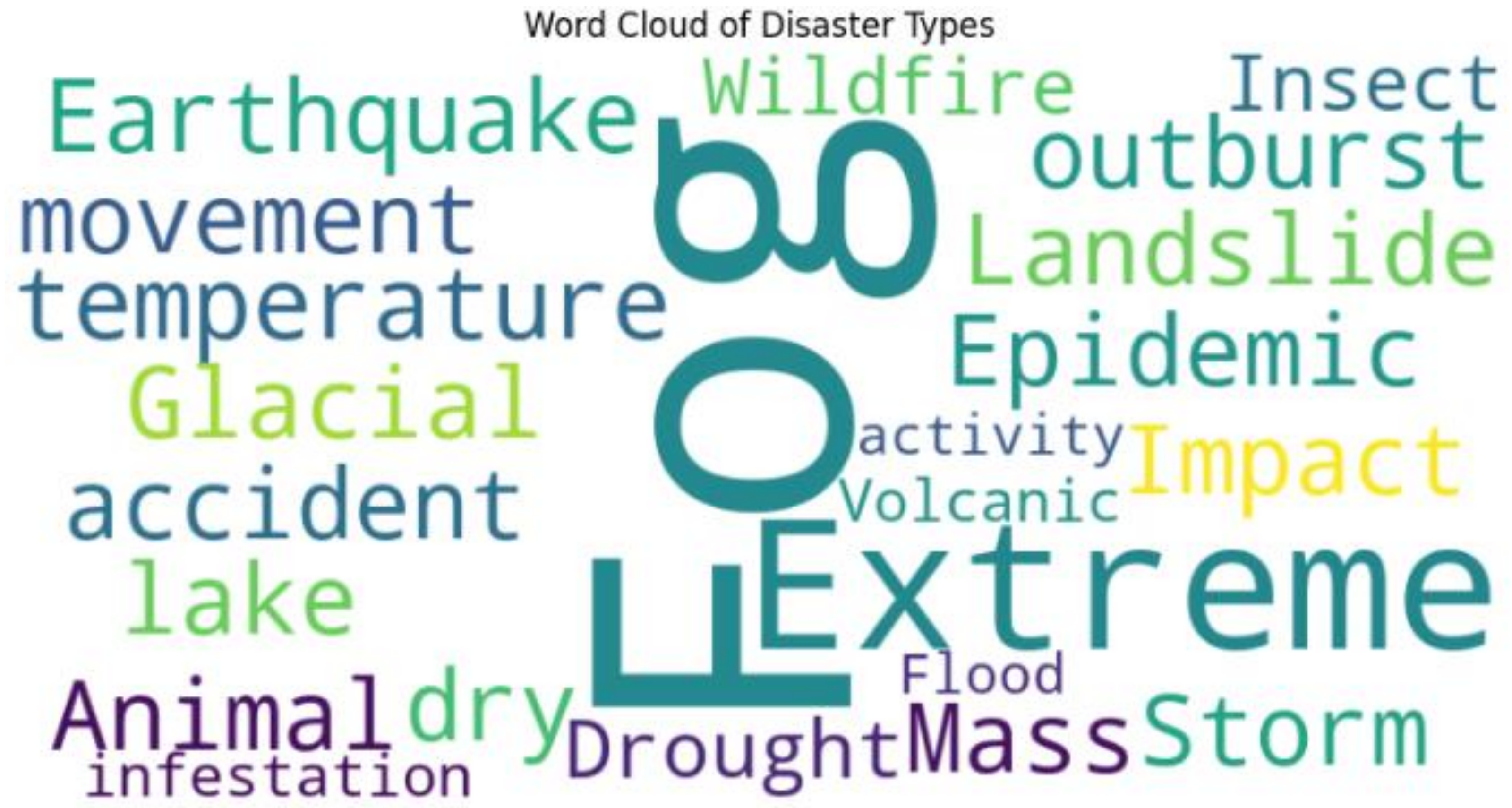
Comparison of Number of Homeless People between Floods and Earthquakes
1e8
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
No Homeless
Earthquake
Flood
Disaster Type

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 Flooc

# 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.
```


Distribution of Disaster Subgroups
Extra-terrestrial
Climatological
Biological
Geophysical
Meteorological
Hydrological
Disaster Subgroup
0.0%
7.2%
8.8%
12.4%
30.0%
41.7%

- **Frequency of Types of Disasters:**

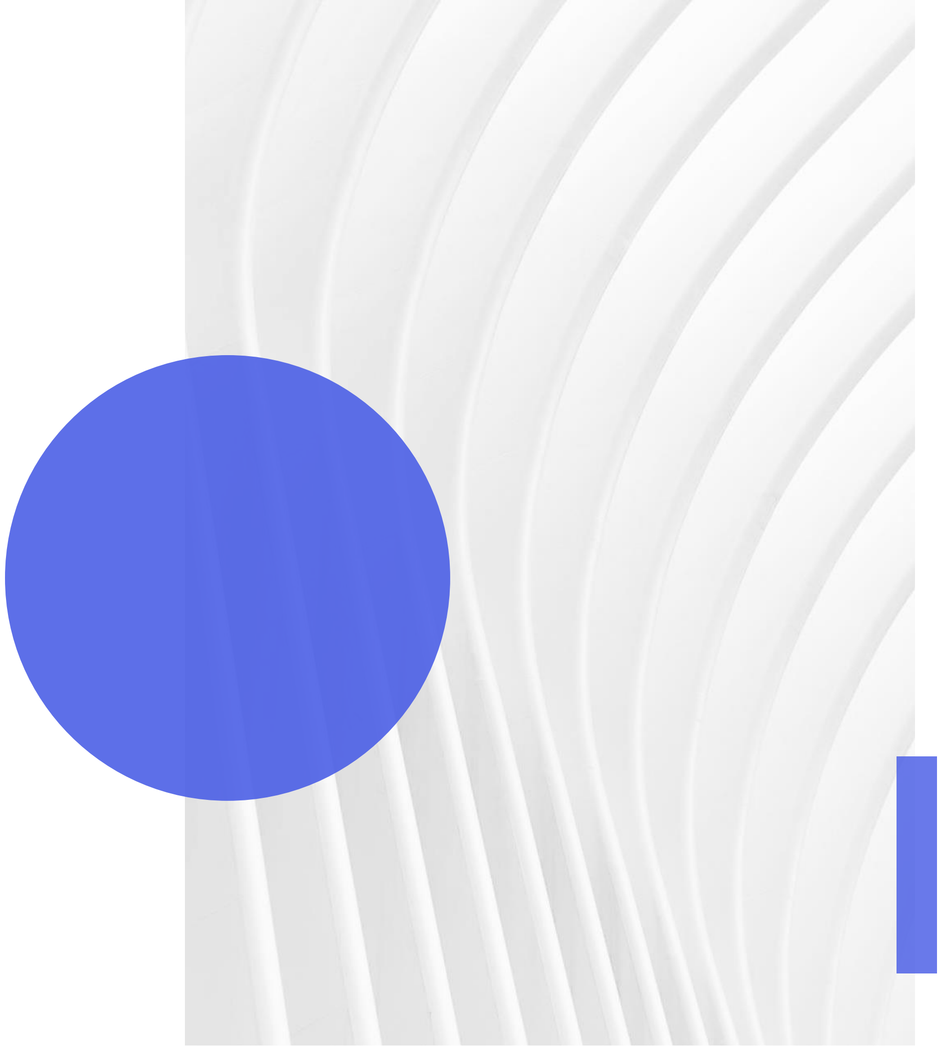


CONCLUSION

- 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.



THANK YOU