PROJECT FILE ON NATURAL DISASTER DATA ANALYSES

GROUP 1

SUBMITTED TO : MR. CHINTAN PATEL | SUBMITTED BY: AKSHIT PATHAK | ANSHUL CHATURVEDI | SAKET KUMAR | SHUBHAMKUMAR | ABHISHEK KUMAR | AMIT RANJAN

INTRODUCTION

In an era marked by increasing environmental challenges and anthropogenic activities, understanding the dynamics of disaster events is crucial for informed decision-making and effective disaster management. The "Disaster Events Dataset" provides a comprehensive and detailed repository of information on a diverse array of disasters that have unfolded globally. This dataset, characterized by a rich set of metadata, aims to facilitate a nuanced exploration of various disaster dimensions, ranging from the humanitarian impact on populations to the economic repercussions on affected regions.

The dataset encompasses a multitude of key parameters, each shedding light on different facets of disaster events. From the overarching categorization into Disaster Category and Disaster Type to the granular details of Event Name, Location, and Origin, the dataset offers a multi-dimensional perspective. Moreover, temporal aspects such as the start and end dates provide a chronological context, enabling the analysis of patterns and trends over time.

Humanitarian efforts and aid contributions are crucial in the aftermath of disasters. The inclusion of fields such as AID Contribution reflects the global community's response to these events, while metrics like Total Deaths, No. Injured, and No. Homeless provide insights into the immediate impact on human lives. Additionally, economic ramifications are captured through metrics such as Reconstruction Costs, Insured Damage, and Total Damage, contributing to a holistic understanding of the overall consequences.

To ensure the dataset's relevance to contemporary economic conditions, the Consumer Price Index (CPI) has been incorporated, allowing for the adjustment of economic metrics to real-world inflationary trends.

This introduction sets the stage for delving into the wealth of information contained in the Disaster Events Dataset, emphasizing its significance in comprehending the diverse nature of disasters, their implications, and the collective responses that shape our global resilience in the face of adversity.

In [7]:

```
#LIBRARY USED IN THIS PROJECT
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [8]:

```
#DATASET
df = pd.read_excel('GROUP_ET_PFM_2.xlsx')
df
```

Out[8]:

	Disaster Category	Disaster Type	Disaster Subtype	Event Name	Country	Subregion	Region	
0	Meteorological	Storm	Tropical cyclone	NaN	United States of America	Northern America	Americas	
1	Hydrological	Flood	Flood (General)	NaN	Jamaica	Latin America and the Caribbean	Americas	
2	Biological	Epidemic	Viral disease	Gastroenteritis	Jamaica	Latin America and the Caribbean	Americas	
3	Geophysical	Volcanic activity	Ash fall	NaN	Japan	Eastern Asia	Asia	
4	Geophysical	Earthquake	Ground movement	NaN	Türkiye	Western Asia	Asia	KARS
...	...	...	...	...	...	...	...	
16862	Meteorological	Storm	Severe weather	NaN	India	Southern Asia	Asia	
16863	Hydrological	Flood	Flood (General)	NaN	Australia	Australia and New Zealand	Oceania	
16864	Climatological	Drought	Drought	NaN	Honduras	Latin America and the Caribbean	Americas	
16865	Climatological	Drought	Drought	NaN	Spain	Southern Europe	Europe	
16866	Climatological	Drought	Drought	NaN	Indonesia	South-eastern Asia	Asia	Lar

16867 rows × 30 columns

In [9]:
```
#META DATA
df.columns
```

Out[9]:
```
Index(['Disaster Category', 'Disaster Type', 'Disaster Subtype', 'Event Nam
e',
       'Country', 'Subregion', 'Region', 'Location', 'Origin',
       'Associated Types', 'AID Contribution', 'Magnitude', 'Magnitude Scal
e',
       'Start Year', 'Start Month', 'Start Day', 'End Year', 'End Month',
       'End Day', 'Total Deaths', 'No. Injured', 'No. Affected',
       'No. Homeless', 'Total Affected', 'Reconstruction Costs',
       'Insured Damage', 'Total Damage', 'CPI', 'Entry Date', 'Last Updat
e'],
      dtype='object')
```

Before getting into data analytics let us try to find is there any null values or missing data in this data set and find the summary statistics of the data to get an idea about the dataset.

In [10]:
```
df.update(df[['Total Deaths','No. Injured','No. Affected','No. Homeless','Tota
              'Reconstruction Costs','Insured Damage','Total Damage','CPI']].f
```

Here, to treat null values, a concept of Constant Value Imputation is used. As 'Total Deaths','No. Injured','No. Affected','No. Homeless','Total Affected', 'Reconstruction Costs','Insured Damage','Total Damage','CPI' are the absolute values and null here means zero. Therefore null values are been filled with 0.

In [11]:
```
df.drop(columns=['Origin','Associated Types','AID Contribution','Magnitude'],
```

Columns which have no significance in the future analyses are dropped from the data set

In [13]:
```
df.isnull().sum()
```

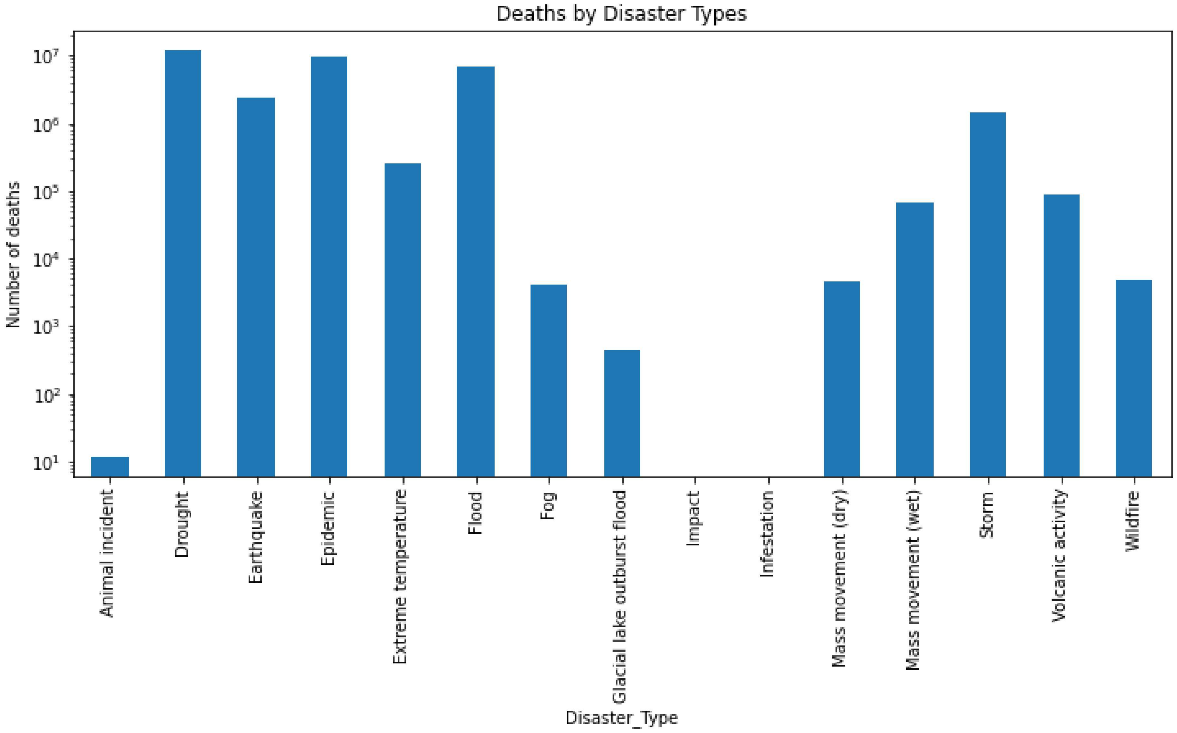
Out[13]:
```
Disaster Category          0
Disaster Type              0
Disaster Subtype           0
Event Name             12849
Country                    0
Subregion                  0
Region                     0
Location                1858
Magnitude Scale         2019
Start Year                 0
Start Month              396
Start Day               3625
End Year                   0
End Month                700
End Day                 3538
Total Deaths               0
No. Injured                0
No. Affected               0
No. Homeless               0
Total Affected             0
Reconstruction Costs       0
Insured Damage             0
Total Damage               0
CPI                        0
Entry Date                 0
Last Update                0
dtype: int64
```

FINDINGS

1. Number of total deaths with respect to the disaster type

In [15]:
```
df.groupby("Disaster Type")["Total Deaths"].sum().plot(figsize=(12,5),
kind="bar", xlabel="Disaster_Type", ylabel="Number of deaths", title="Deaths b
```

Out[15]:
```
<AxesSubplot:title={'center':'Deaths by Disaster Types'}, xlabel='Disaster_T
ype', ylabel='Number of deaths'>
```



Interpretation: Here we can see that the most number of deaths has been due to Drought. The govenment should implement policies to deal with this problem.

2. Number of total deaths with respect to the Subregion

In [25]:

```
df.groupby('Subregion')['Total Deaths'].sum().sort_values()
```

Out[25]:

```
Subregion
Micronesia                              128.0
Polynesia                               769.0
Central Asia                           3566.0
Australia and New Zealand              9744.0
Melanesia                             12162.0
Northern Europe                       13632.0
Western Europe                        66269.0
Northern America                      97631.0
Western Asia                         163785.0
Southern Europe                      205408.0
Northern Africa                      220279.0
South-eastern Asia                   511101.0
Latin America and the Caribbean      738741.0
Sub-Saharan Africa                  1245458.0
Eastern Europe                      3936119.0
Southern Asia                      12570677.0
Eastern Asia                       12817165.0
Name: Total Deaths, dtype: float64
```

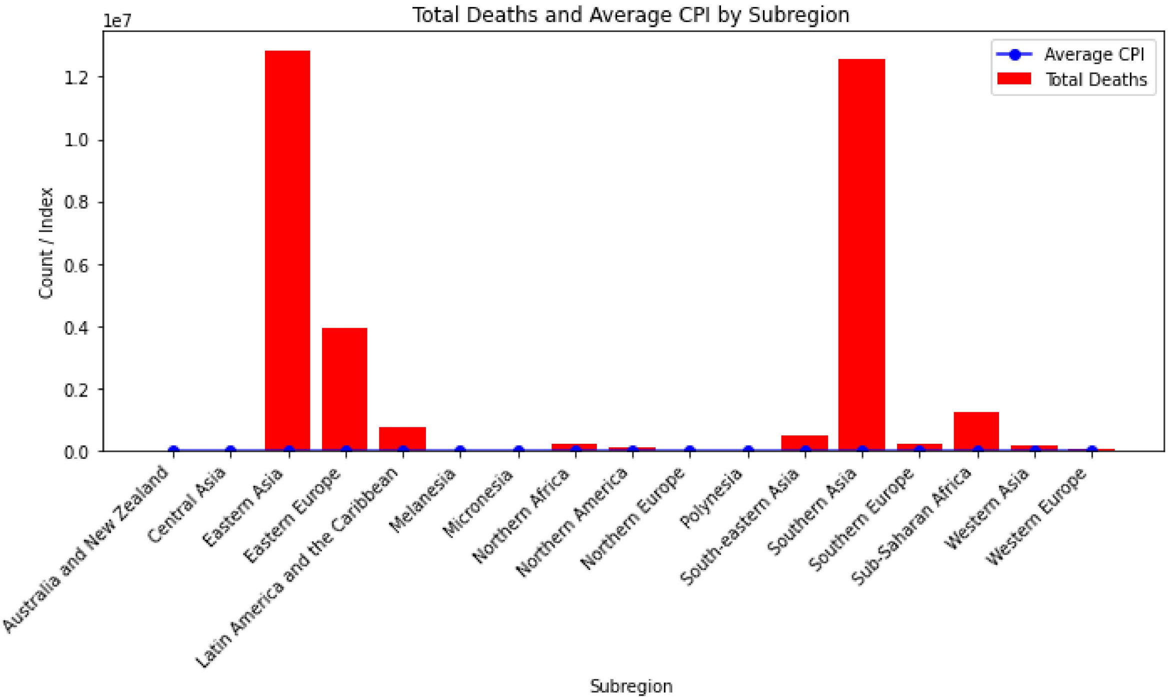
Interpretation: Here we can see that the most number of deaths has been seen in Eastern Asia, while the lowest number of deaths has been shown in Micronesia.

3. Relaionship between CPI and Total number of deaths, grouped by Subregion

In [26]:

```
import pandas as pd
import matplotlib.pyplot as plt
result = df.groupby('Subregion').agg({'Total Deaths': 'sum', 'CPI': 'mean'}).r

print(result)

plt.figure(figsize=(10, 6))
plt.bar(result['Subregion'], result['Total Deaths'], label='Total Deaths', col
plt.plot(result['Subregion'], result['CPI'], label='Average CPI', color='blue'
plt.xlabel('Subregion')
plt.ylabel('Count / Index')
plt.title('Total Deaths and Average CPI by Subregion')
plt.legend()
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

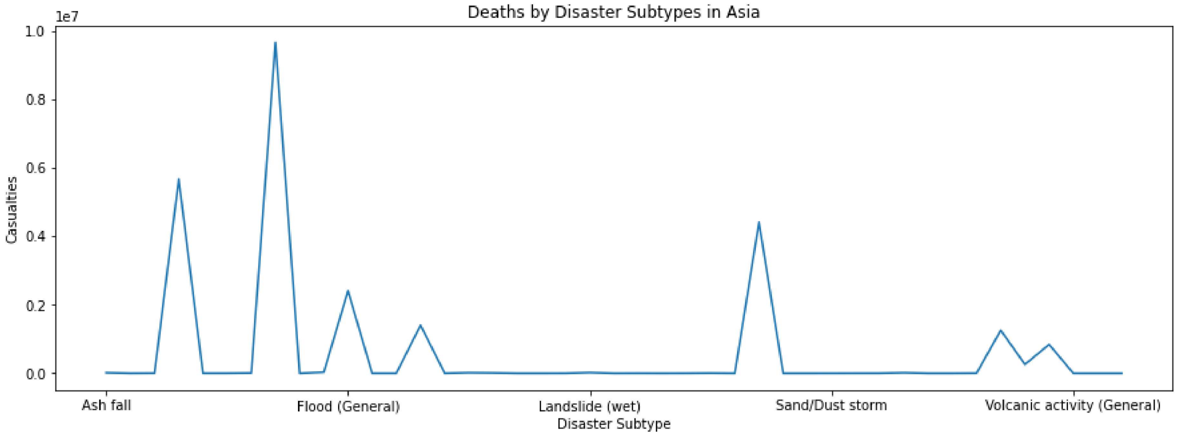
```
                          Subregion  Total Deaths        CPI
0         Australia and New Zealand        9744.0  55.818554
1                      Central Asia        3566.0  66.267531
2                      Eastern Asia    12817165.0  54.272361
3                    Eastern Europe     3936119.0  60.164798
4   Latin America and the Caribbean      738741.0  55.927633
5                         Melanesia       12162.0  52.996433
6                        Micronesia         128.0  63.231171
7                   Northern Africa      220279.0  54.727318
8                  Northern America       97631.0  54.652882
9                   Northern Europe       13632.0  61.679439
10                        Polynesia         769.0  49.193736
11               South-eastern Asia      511101.0  59.062901
12                    Southern Asia    12570677.0  54.318373
13                  Southern Europe      205408.0  56.227759
14               Sub-Saharan Africa     1245458.0  62.531142
15                     Western Asia      163785.0  54.027442
16                   Western Europe       66269.0  57.611421
```



Interpretation: There is a negative correlation between the total number of deaths and the average CPI Index. High average CPI have low death numbers, as they are already prepared for any natural disaster.

4. Deaths by Disaster Subtypes in Asia

In [27]:
```
df[df["Region"]=="Asia"].groupby(["Disaster Subtype"])["Total Deaths"].sum().p
logy=False, title="Deaths by Disaster Subtypes in Asia")
plt.show()
```



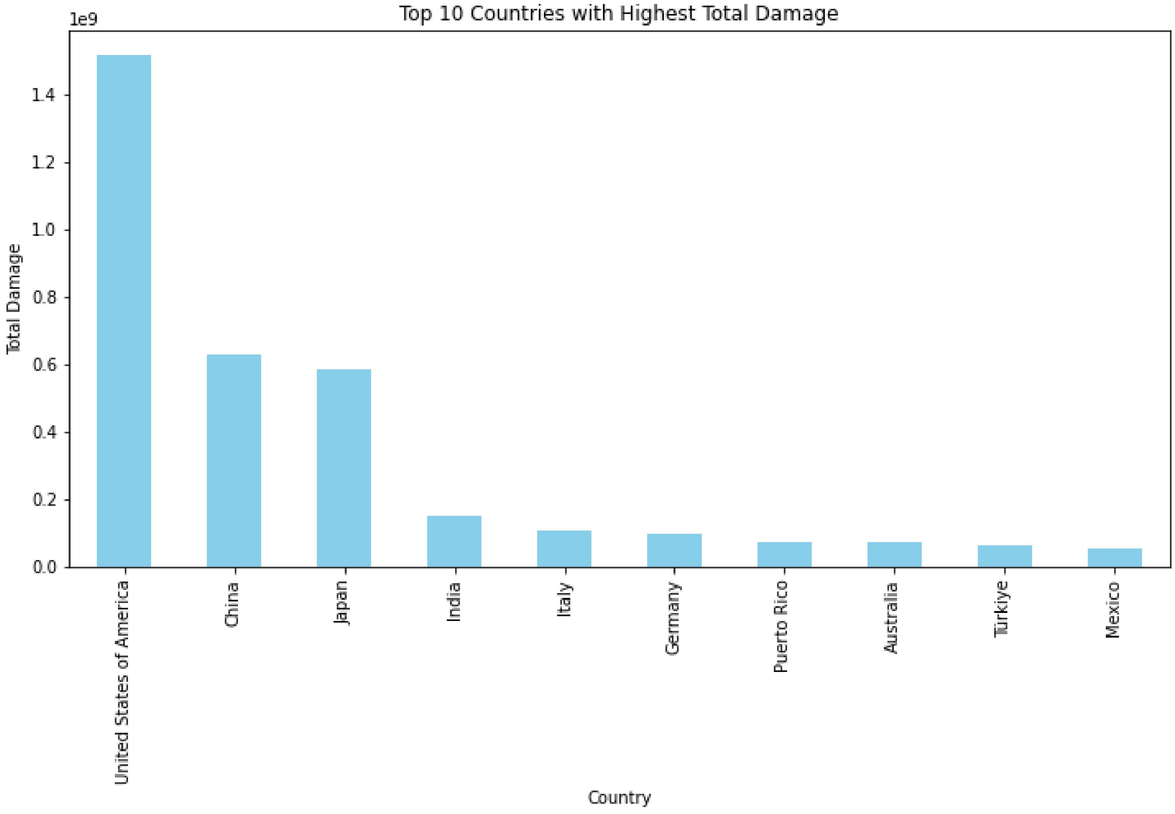
5. Top 5 countries with the highest total affected population

In [23]:
```
top_affected_countries = df.groupby('Country')['Total Affected'].sum().nlarges
print(f"4. Top 5 countries with the highest total affected population:\n{top_a
```

```
4. Top 5 countries with the highest total affected population:
Country
China          3.323233e+09
India          2.483116e+09
Bangladesh     4.730051e+08
Philippines    2.625574e+08
Pakistan       1.322488e+08
Name: Total Affected, dtype: float64
```

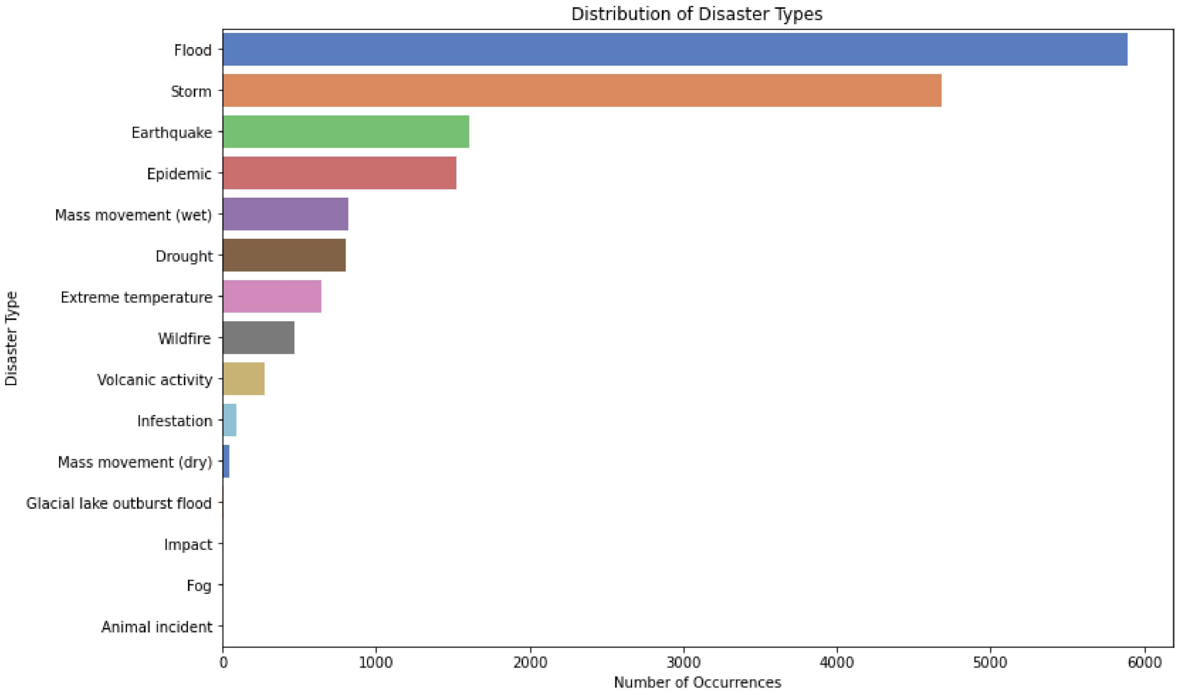
6. Top 10 Countries with Highest Total Damage

In [17]:
```
top_countries_damage = df.groupby('Country')['Total Damage'].sum().nlargest(1
plt.figure(figsize=(12, 6))
top_countries_damage.plot(kind='bar', color='skyblue')
plt.title('Top 10 Countries with Highest Total Damage')
plt.xlabel('Country')
plt.ylabel('Total Damage')
plt.show()
```



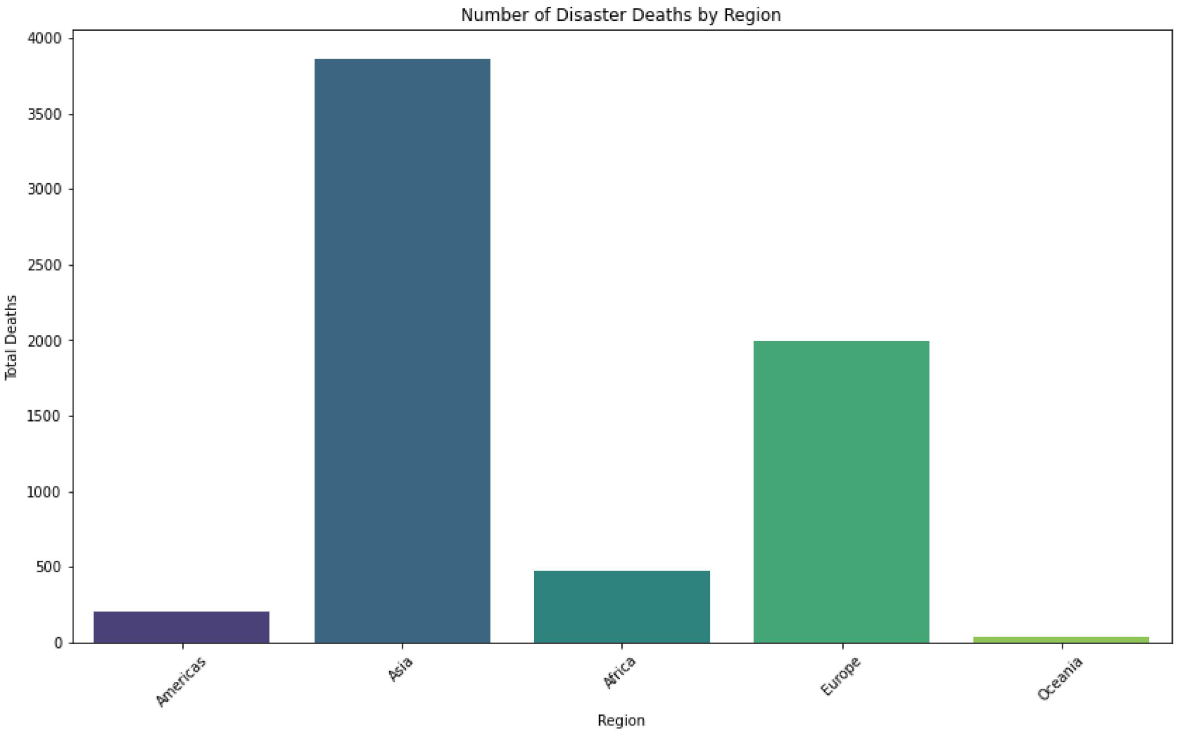
7. Distribution of Disaster Types

In [30]:
```
plt.figure(figsize=(12, 8))
sns.countplot(y='Disaster Type', data=df, order=df['Disaster Type'].value_cour
plt.title('Distribution of Disaster Types')
plt.xlabel('Number of Occurrences')
plt.ylabel('Disaster Type')
plt.show()
```



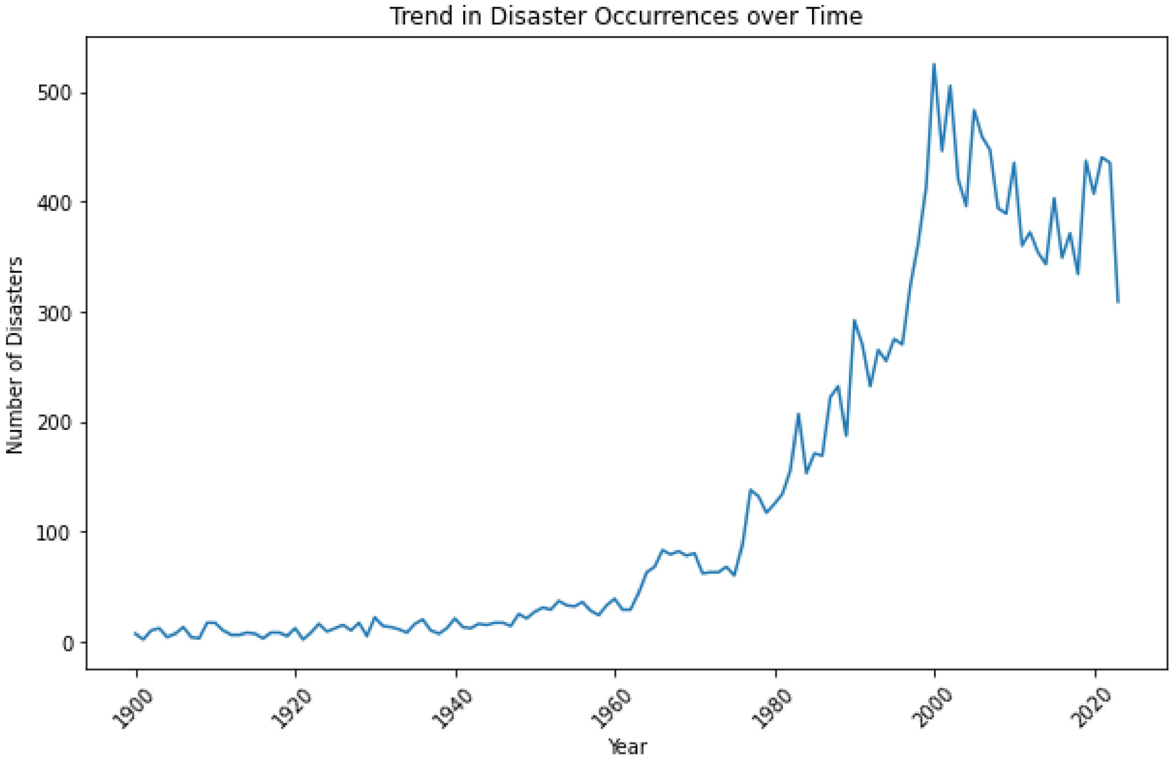
8. Number of Disaster Deaths by Region

In [32]:
```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(14, 8))
sns.barplot(x='Region', y='Total Deaths', data=df, palette='viridis', ci=None)
plt.title('Number of Disaster Deaths by Region')
plt.xlabel('Region')
plt.ylabel('Total Deaths')
plt.xticks(rotation=45)
plt.show()
```



9. Trend in Disaster Occurrences over Time

In [31]:
```
disasters_per_year = df.groupby('Start Year').size()

plt.figure(figsize=(10, 6))
disasters_per_year.plot(kind='line')
plt.xlabel('Year')
plt.ylabel('Number of Disasters')
plt.title('Trend in Disaster Occurrences over Time')
plt.xticks(rotation=45)
plt.show()
```



Conclusion:

In conclusion, the dataset provides a comprehensive overview of various disasters, encompassing factors such as disaster category, type, subtype, event name, country, subregion, region, location, origin, associated types, aid contributions, magnitude, magnitude scale, and temporal details including start and end dates. The dataset also includes impact metrics such as total deaths, number injured, number affected, number homeless, total affected, and financial measures such as reconstruction costs, insured damage, and total damage. Additionally, the Consumer Price Index (CPI) is included, although its specific meaning or context within this dataset requires further clarification.

Overall, this dataset serves as a valuable resource for disaster research and emergency management, offering a multidimensional perspective that combines geographical, temporal, and impact-related aspects. To enhance the dataset's utility, it is recommended to seek clarification on the meaning of ambiguous variables, explore potential correlations with external factors, and consider the dataset's limitations in drawing comprehensive conclusions.