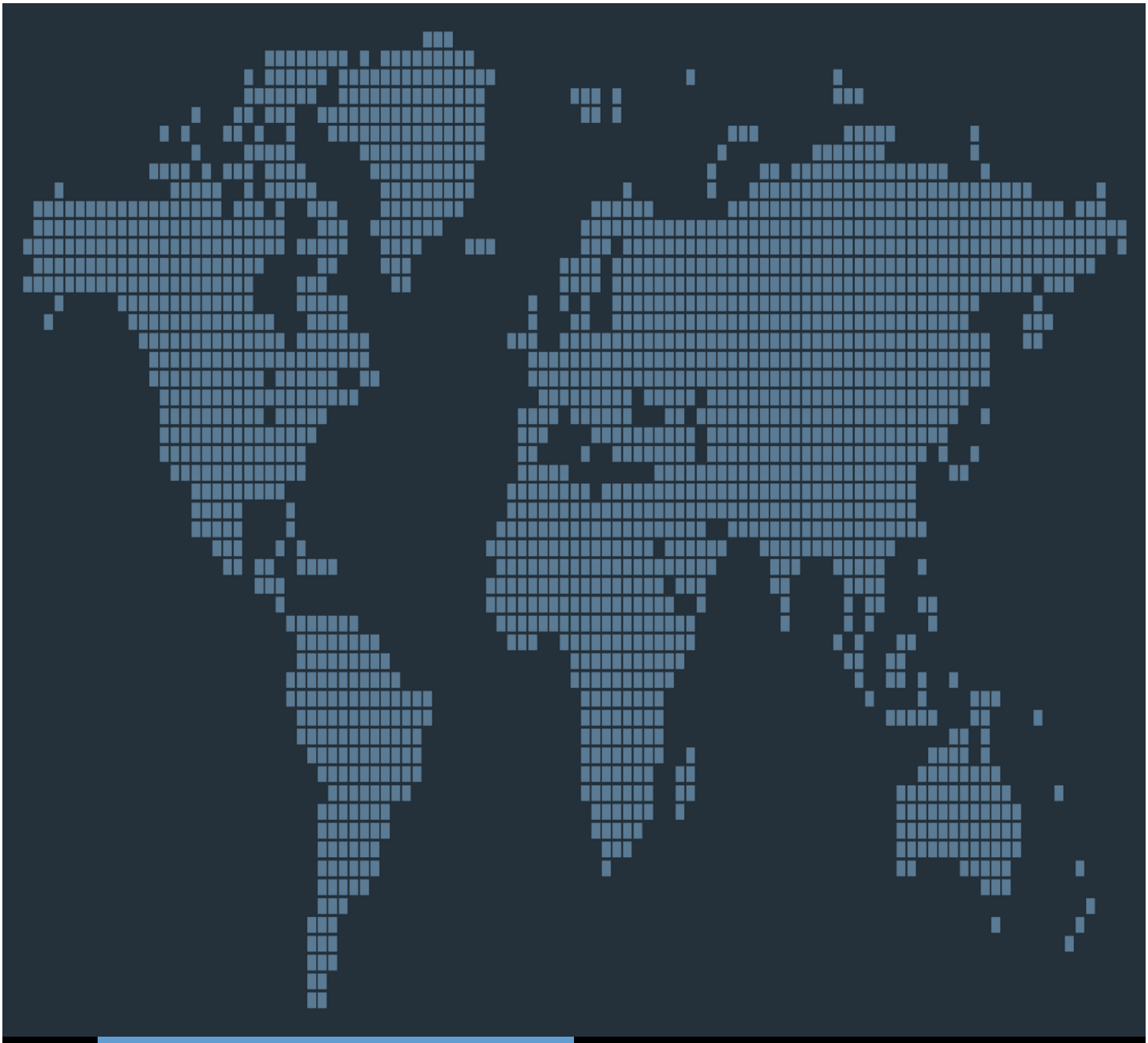


GLOBAL TERRORISM DATABASE

DATA ANALYSIS

AHMED OSAMA AHMED ABDELAAL

INTRODUCTION

This report analyzes the Global Terrorism Database (GTD), an open-source collection of over 180,000 terrorist incidents worldwide from 1970 through 2017, maintained by the START consortium at the University of Maryland. It details domestic and international attacks, providing a foundation for exploring trends and impacts of terrorism. The analysis aims to identify patterns, affected regions, attack types and targets.

DATA PREPROCESSING

Renaming and Select Relevant Columns:

Out of the original **135** columns in the dataset, **18** columns were identified as most relevant, and the selected columns were also **renamed** for clarity and ease of use.

```
Data columns (total 18 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Year         181691 non-null  int64
 1   Month        181691 non-null  int64
 2   Day          181691 non-null  int64
 3   Country      181691 non-null  object
 4   Region       181691 non-null  object
 5   city         181256 non-null  object
 6   latitude     177135 non-null  float64
 7   longitude    177134 non-null  float64
 8   AttackType   181691 non-null  object
 9   Killed       171378 non-null  float64
 10  Wounded      165380 non-null  float64
 11  Target       181053 non-null  object
 12  Summary      115562 non-null  object
 13  Group        181691 non-null  object
 14  Target_type  181691 non-null  object
 15  Weapon_type  181691 non-null  object
 16  Motive       50561 non-null   object
 17  casualities  164817 non-null  float64
dtypes: float64(5), int64(3), object(10)
```

Day and Month Values:

It was observed that **891 records** had a **'Day'** value of **0**, and **20 records** had a **'Month'** value of **0**, these records were removed.

Handle Missing Values:

The **'Motive'** and **'Summary'** columns, with missing data percentages of **72.17%** and **36.39%** respectively, were removed from the dataset. This decision was based on the high proportion of missing entries, and **'Summary'** is not very relevant to the Analysis.
After removing the columns, the **removal** of remaining records with **null values** was performed because they were so low compared to the dataset's size.

```
Year            0
Month           0
Day             0
Country         0
Region          0
city          435
latitude     4556
longitude    4557
AttackType      0
Killed      10313
Wounded     16311
Target        638
Summary     66129
Group           0
Target_type     0
Weapon_type     0
Motive     131130
casualities 16874
dtype: int64
```

```
Year           0.000000
Month          0.000000
Day            0.000000
Country        0.000000
Region         0.000000
city           0.239417
latitude       2.507554
longitude      2.508104
AttackType     0.000000
Killed         5.676120
Wounded        8.977330
Target         0.351146
Summary       36.396409
Group          0.000000
Target_type    0.000000
Weapon_type    0.000000
Motive        72.171984
casualities    9.287196
dtype: float64
```

```
Year          0
Month         0
Day           0
Country       0
Region        0
city          0
latitude      0
longitude     0
AttackType    0
Killed        0
Wounded       0
Target        0
Summary       0
Group         0
Target_type   0
Weapon_type   0
casualities   0
dtype: int64
```

Duplicates Values:

There were 8,897 duplicate records, which was 5.58% of the records, these duplicates were removed.

After Cleaning:

Originally containing **181,691 records**, the dataset now consists of **150,421 records**. This reduction reflects the **removal of 17%** of the original data. A **17%** reduction in a dataset of this size is of course **significant**.

DATA ANALYSIS

Descriptive Statistics:

```
Mean killed: 2.1933174224343674, Median killed: 0.0, Std killed: 10.032976760174048
Mean wounded: 3.3577293064133333, Median wounded: 0.0, Std wounded: 37.628016496197134
```

Killed:

- **Mean (Average):** The average number of individuals killed per attack **2.19**.
- **Median:** The median number of individuals killed is **0.0**, more than half of the recorded attacks result in no fatalities.
- **Standard Deviation:** The standard deviation for killings is **10.33**, indicating a wide variance in the number of fatalities per attack.

Wounded:

- **Mean (Average):** The average number of individuals wounded per attack is **3.36**. This value is slightly higher than that for fatalities, suggesting that attacks tend to injure more individuals than they kill.
- **Median:** The median number of individuals wounded is also **0.0**, which, similar to the fatalities, shows that a significant number of attacks do not result in injuries.
- **Standard Deviation:** The standard deviation for wounded is high at about **37.63**, reflecting a significant variability in the number of injuries per attack.

```
Most frequent country: Iraq
Most frequent region: Middle East & North Africa
Most frequent attack type: Bombing/Explosion
Most frequent target type: Private Citizens & Property
Most frequent weapon type: Explosives
```

Frequency:

- **Country: Iraq** is the most frequently affected country, has a high concentration of terrorist activities.
- **Region: The Middle East & North Africa region** has the highest incidence of attacks, suggesting regional geopolitical and socio-economic challenges.
- **Attack Type: Bombing/Explosion** is the most method used, indicating its effectiveness and potentially low implementation barriers.
- **Target Type: Private Citizens & Property** are the most frequent targets, indicating terrorism's focus on disrupting civilian life and creating widespread fear.
- **Weapon Type: Explosives** are the most used weapon, consistent with the most frequent type of attacks.

Geographical Impact and Terrorism Groups:

Most Affected Regions:

1. **Middle East & North Africa:** Leads with the highest number of terrorist attacks, indicates the region's significant geopolitical volatility.
2. **South Asia:** Follows closely, shows the ongoing regional conflicts and political instability.

Most Affected Countries:

1. **Iraq:** Most impacted with 21,675 attacks, highlighting it as a central focus of terrorism activities.
2. **Pakistan and Afghanistan**: Also highly affected, with 13,141 and 11,484 attacks, respectively, indicating the South Asian regional security challenges.

Top Terrorism Groups:

1. **Unknown**: Most attacks were not linked to any known group.
2. **Taliban**: The most active known group with **6,616 attacks**, indicating its significant influence in the region.
3. **Islamic State of Iraq and the Levant (ISIL):** Involved in **4,256 attacks**, indicating its major role in Middle East conflicts.

```
Region
Middle East & North Africa     43543
South Asia                     40685
South America                  14176
Sub-Saharan Africa             13302
Western Europe                 12382
Southeast Asia                 10813
Central America & Caribbean     6259
Eastern Europe                  4676
North America                   3143
East Asia                        657
dtype: int64
```

```
Country
Iraq              21675
Pakistan          13141
Afghanistan       11484
India             10887
Colombia           6520
Philippines        6076
Peru               4192
Turkey             3750
Thailand           3384
United Kingdom     3243
dtype: int64
```

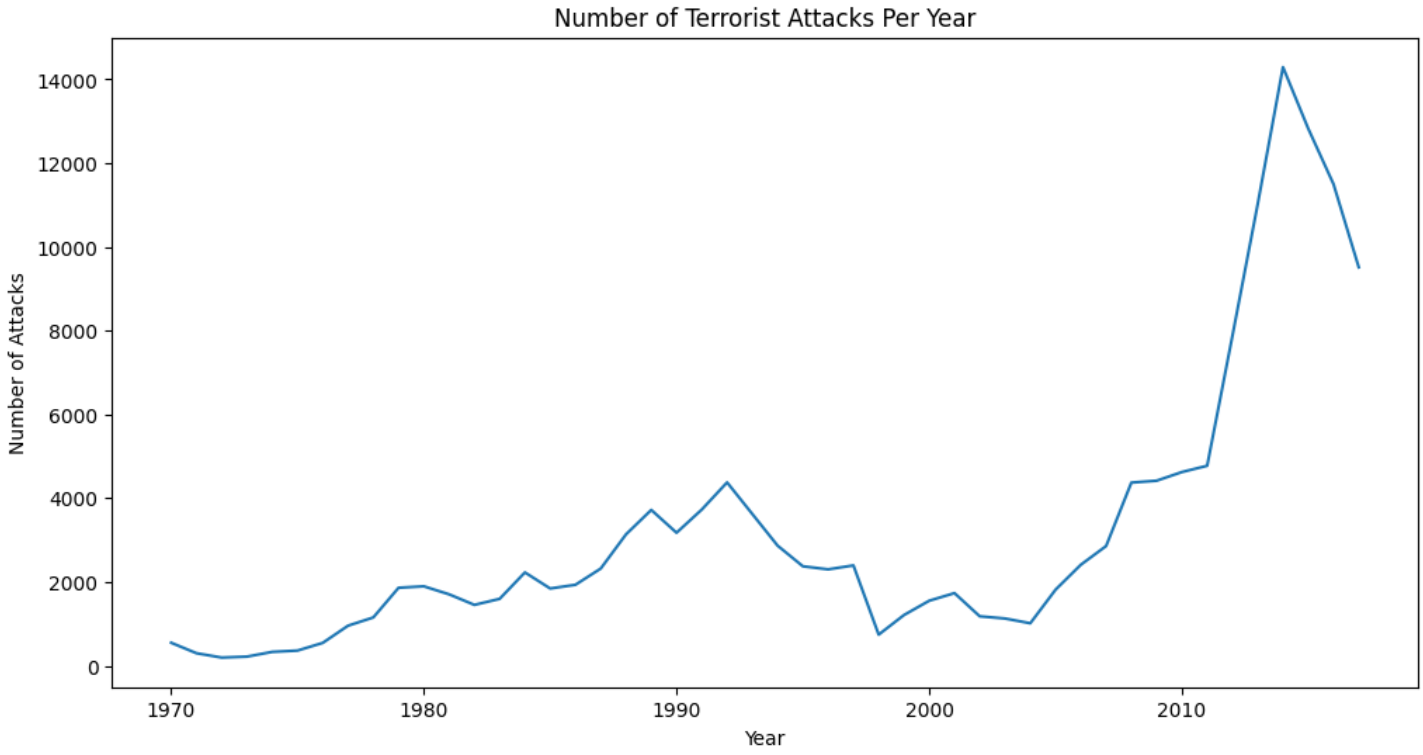
```
Group
Unknown                                    73166
Taliban                                     6616
Islamic State of Iraq and the Levant (ISIL)     4256
Shining Path (SL)                           3084
Al-Shabaab                                  2333
Name: count, dtype: int64
```

VISUALIZATIONS

Trends in Terrorist Attacks Over Time:

Overview: This line chart tracks the number of global terrorist attacks from 1970 to 2017.

Observations: The data shows a general increase over the decades, peaking in 2014 before declining. This rise and fall may correlate with major global and political events.

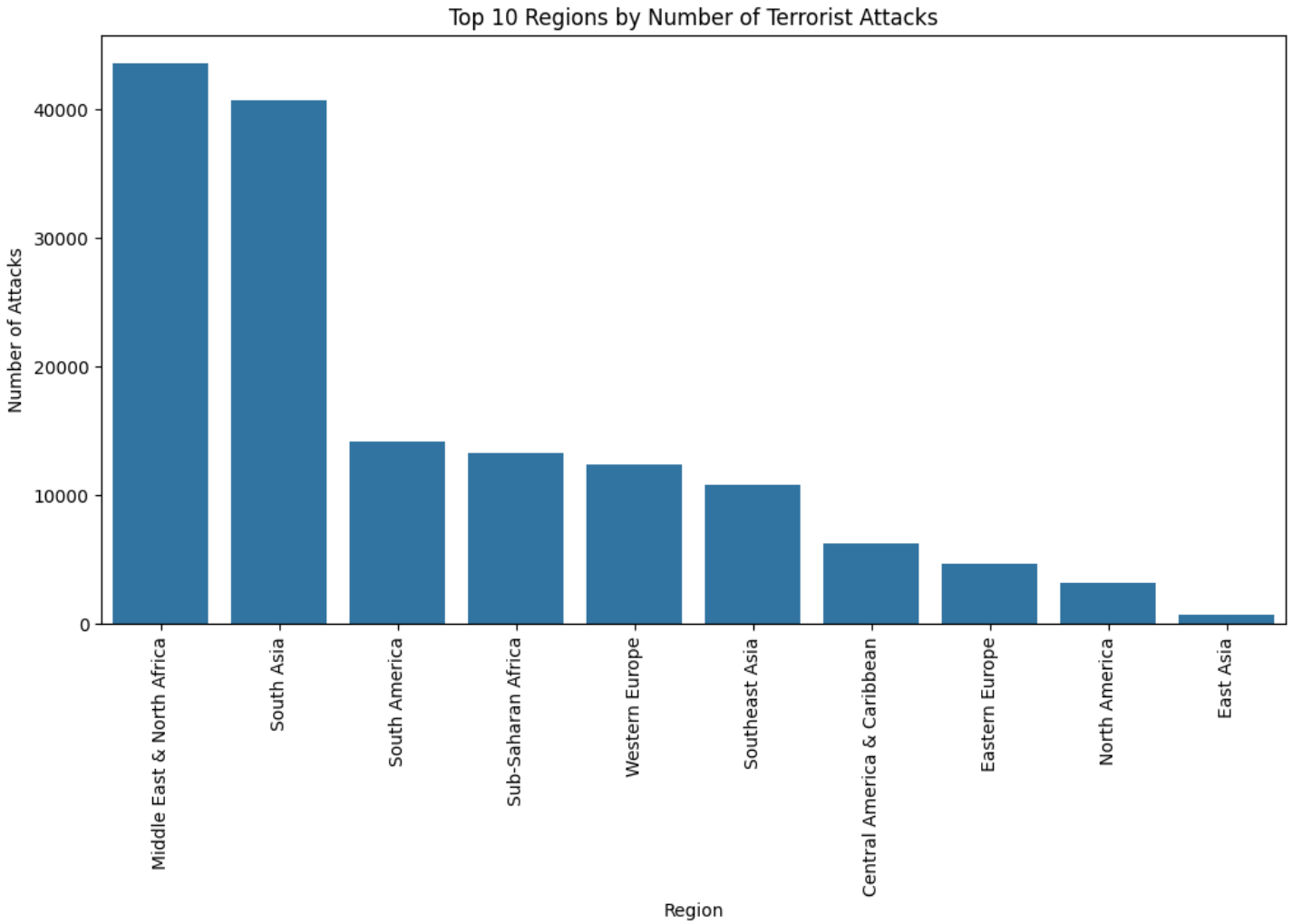


Regional Distribution of Terrorist Attacks:

Overview: The chart ranks the top 10 regions based on the total number of terrorist attacks recorded.

Key Observations:

- **Middle East & North Africa** and **South Asia** dominate the chart, reflecting their high levels of terrorism-related incidents. These regions significantly outpace others, indicating a concentrated presence of terrorism.
- Regions like **Sub-Saharan Africa** and **Southeast Asia** also show high numbers, but less than half of those in the leading regions.

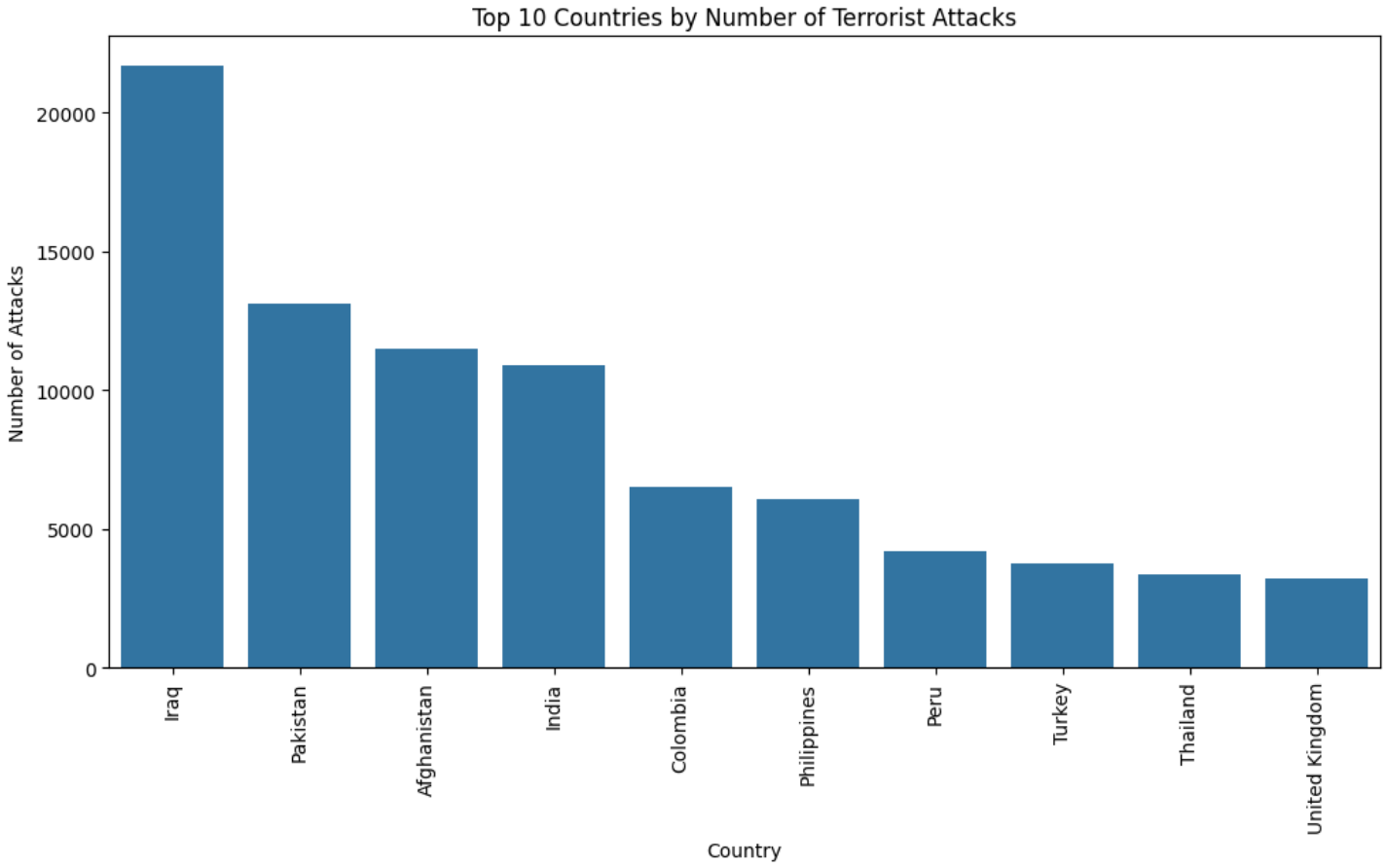


Country-Specific Trends in Terrorist Attacks:

Overview: This chart presents the top 10 countries with the highest numbers of terrorist attacks.

Key Observations:

- **Iraq** leads significantly, highlighting its major challenges with terrorism. It is followed by **Pakistan**, **Afghanistan**, and **India**, all of which also face substantial terrorism incidents.
- Countries like **Colombia**, **Philippines**, **Peru**, **Turkey**, **Thailand**, and **United Kingdom** show comparatively lower frequencies of attacks.

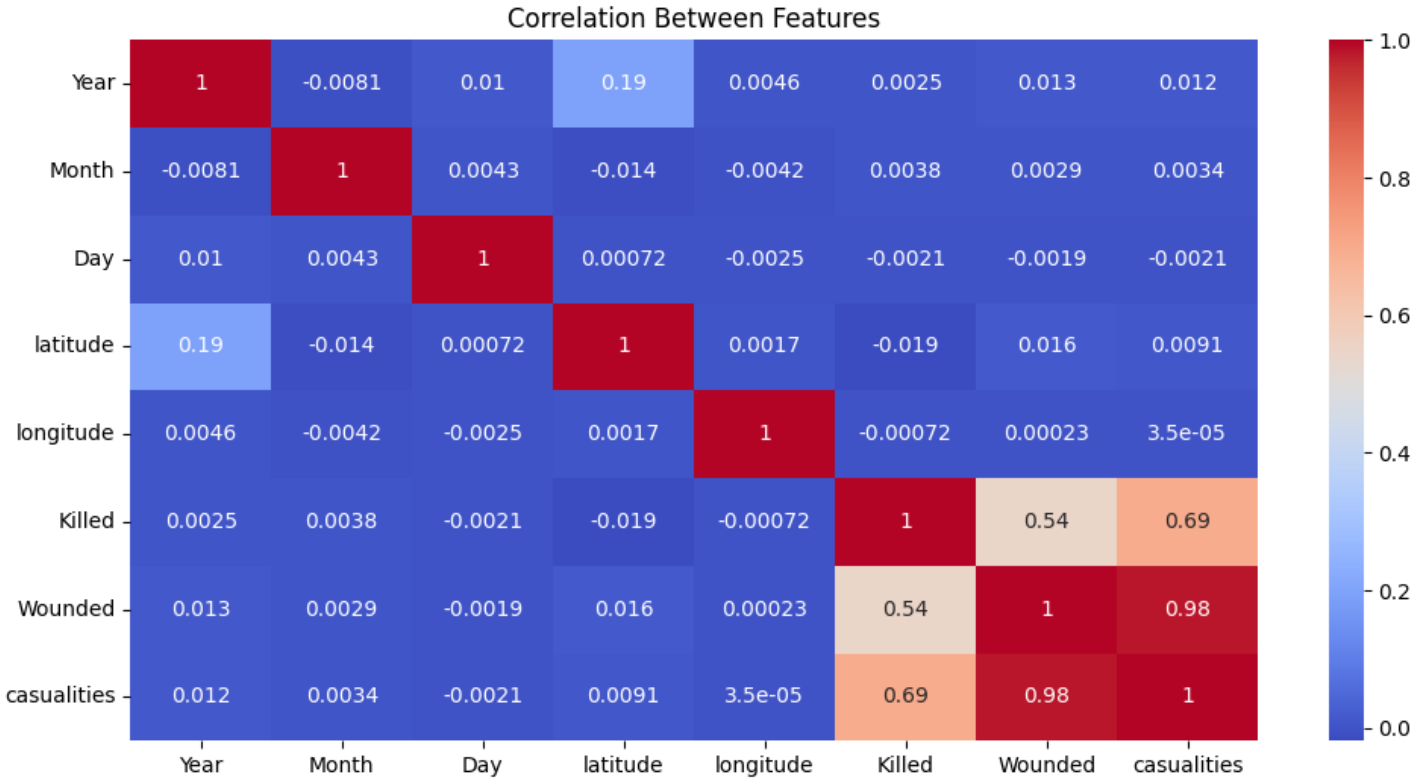


Correlation Heatmap:

Overview: This heatmap displays the correlation between numeric features in the dataset.

Key Observations:

The correlation heatmap showed **no significant** correlations between **features** against **casualty** data, indicating these variables do not linearly influence one another.
Expectedly, **high correlations** were observed among **'Killed'**, **'Wounded'**, and **'Casualties'** is expected.

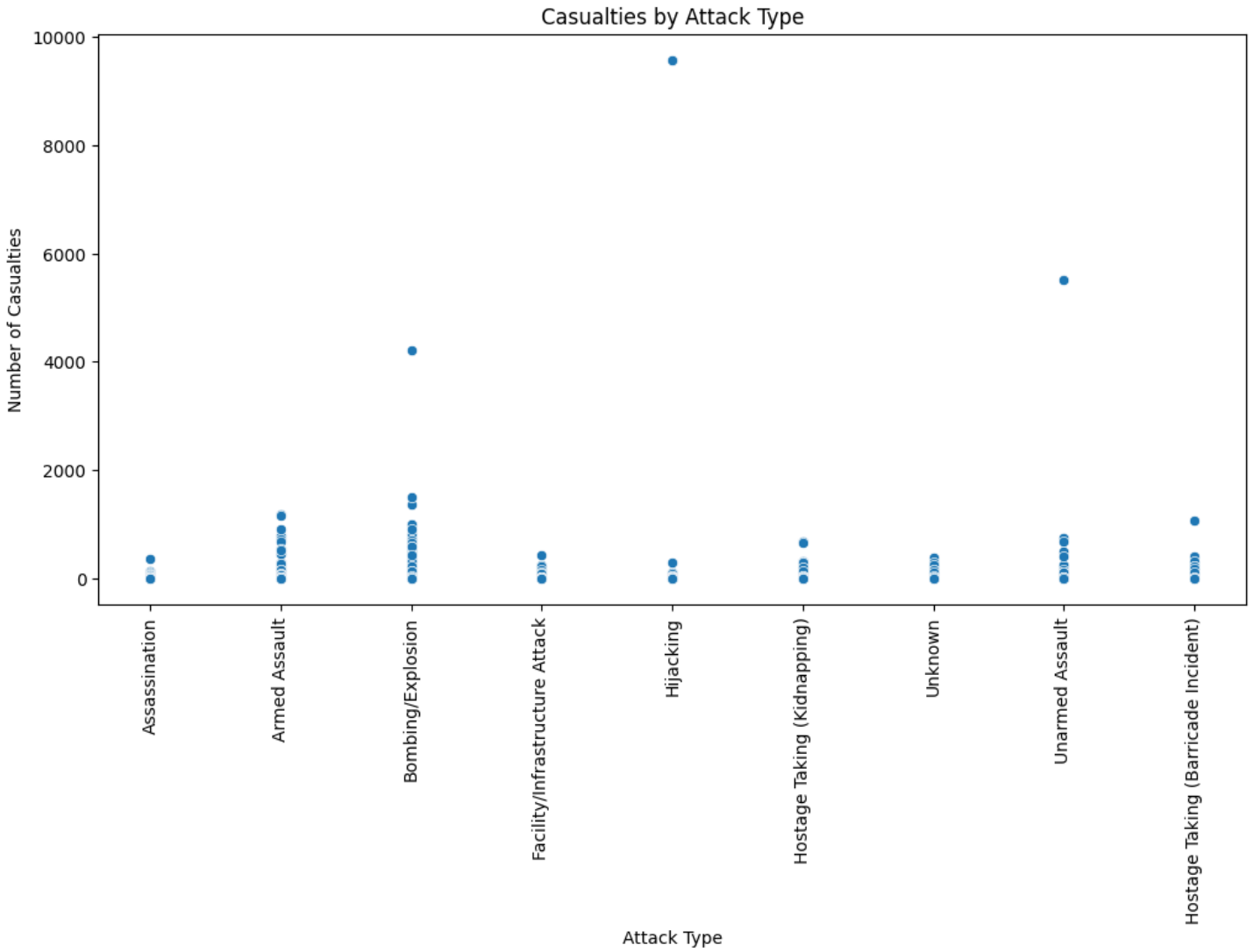


Casualties by Attack Type:

Overview: This scatter plot displays the number of casualties associated with types of terrorist attacks.

Key Observations:

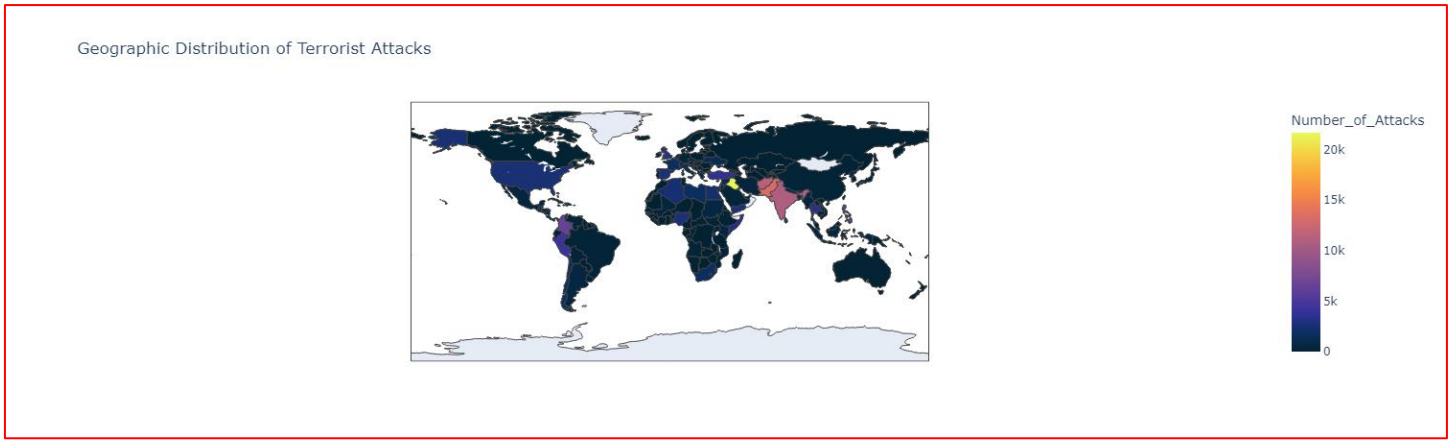
- **Bombing/Explosion** incidents tend to result in the **highest** number of casualties, with several incidents reaching over **4,000 casualties**. This attack type shows not only a high frequency of occurrences but also a high potential for mass casualties.
- **Hostage Taking (Kidnapping)** and **Unarmed Assault** show considerably fewer casualties per incident.

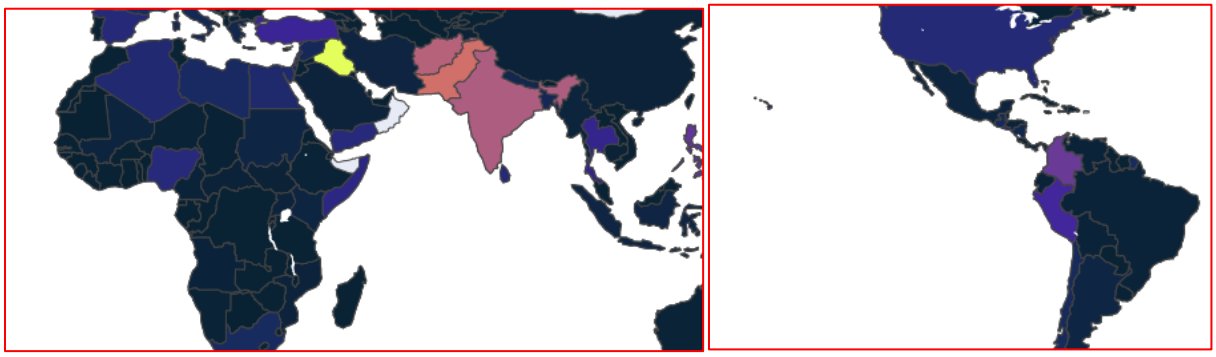


Geographic Distribution of Terrorist Attacks:

Overview: The map displays the number of terrorist attacks by country, shaded according to the scale of attack frequency from 1970 to 2017.
Key Observations:

- **High Concentration Areas:** Regions with the hottest shades, such as the Middle East, North Africa, and South Asia, indicate the highest frequency of terrorist attacks, indicating ongoing regional conflicts.
- **Moderate to Low Concentration:** Countries in North America, Europe, and parts of East Asia show lighter shades, indicating fewer incidents.



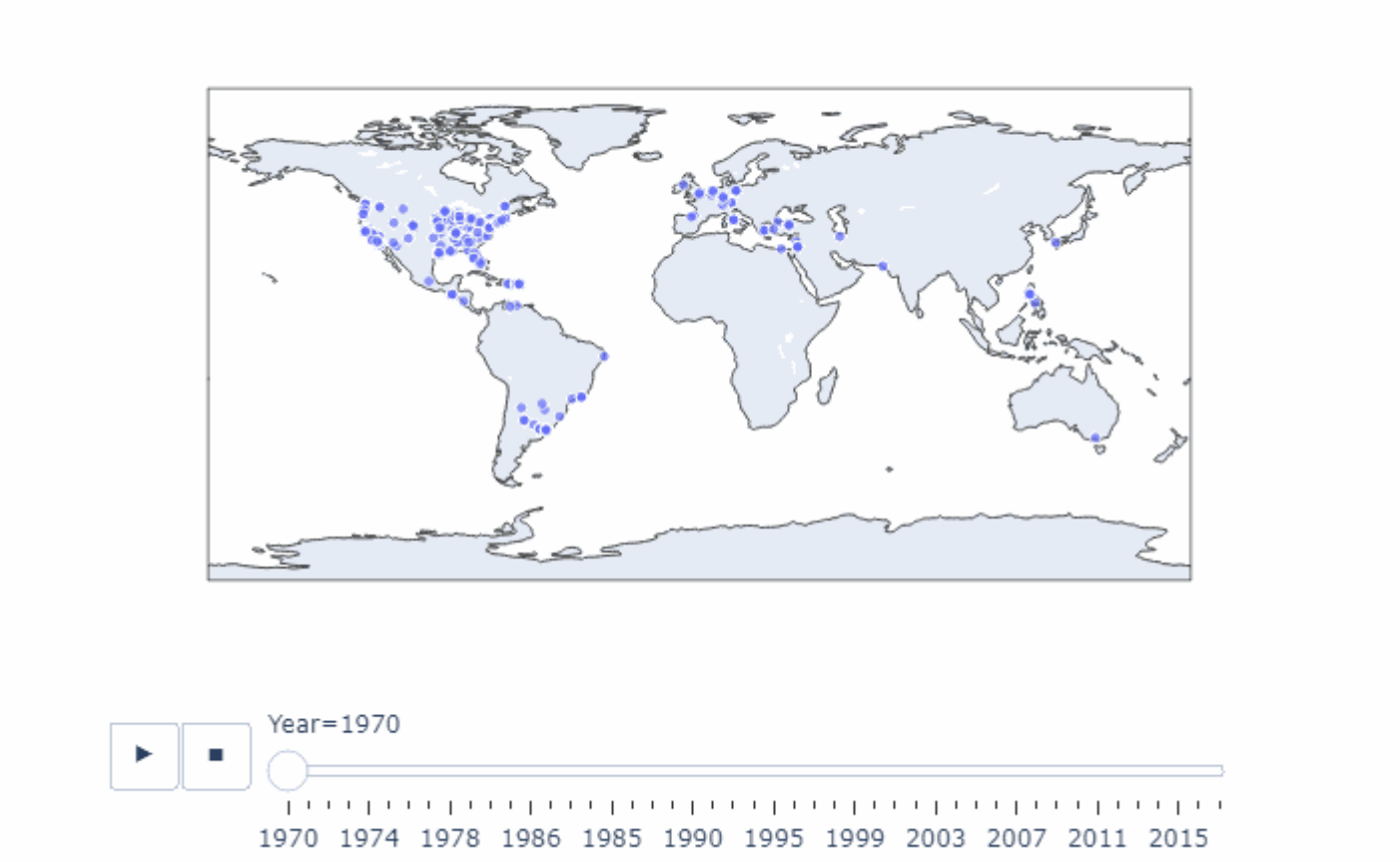


Temporal and Geographic Spread of Terrorist Attacks:

Overview: This animated map illustrates the progression of terrorist attacks worldwide, marking each incident with a dot on the map for each year from 1970 to 2017.

Key Observations:

- **Expansion Over Time:** The density and spread of dots increase over the years, especially in the Middle East, South Asia, and Africa.
- **Shifts in Hotspots:** Initially concentrated in certain areas, such as Western Europe and Latin America in the early decades, the focus shifts towards the Middle East and Africa in later years, reflecting changes in geopolitical situations and the emergence of new terrorist groups.



Performance Comparison with Dask:

```
dask_terror = dd.read_csv('terror_database_clean.csv', encoding='ISO-8859-1')
dask_terror = dask_terror.persist()

def pandas_groupby():
    terror = pd.read_csv('terror_database_clean.csv', encoding='ISO-8859-1')
    return terror.groupby('Year').size()

def dask_groupby():
    result = dask_terror.groupby('Year').size().compute()
    return result

start_time = time.time()
pandas_groupby()
pandas_duration = time.time() - start_time
pandas_memory = memory_usage((pandas_groupby, ), interval=0.1, timeout=120)

start_time = time.time()
dask_groupby()
dask_duration = time.time() - start_time
dask_memory = memory_usage((dask_groupby, ), interval=0.1, timeout=120)

print(f"Pandas duration: {pandas_duration} seconds")
print(f"Dask duration: {dask_duration} seconds")
print(f"Peak memory usage for Pandas: {max(pandas_memory)} MiB")
print(f"Peak memory usage for Dask: {max(dask_memory)} MiB")
```

✓ 7.3s

```
Pandas duration: 0.4500010013580322 seconds
Dask duration: 0.021999597549438477 seconds
Peak memory usage for Pandas: 823.71875 MiB
Peak memory usage for Dask: 774.50390625 MiB
```

Discussion:
After trying different parameters and methods on the dataset, **Dask** showed a **slower** performance in this scenario, it is important to note that **Dask's** strengths would be more visible in scenarios involving much larger datasets where data does not fit into memory. For smaller datasets, as showed, **Pandas** remains **highly** efficient and faster due to its simplicity and lack of parallel computations. And the memory usage between **Dask** and **Pandas** is close in this scenario. **Pandas** is well-suited for data that fits into memory, while **Dask** is designed for parallel computing on larger-than-memory datasets

Insights, Challenges and Limitations:

Insights Highlights:

1. **High-Impact Attack Types:**
   - The analysis identified specific attack types, such as **bombings/explosions**, that consistently result in **high casualties**.

2. **Geographic Hotspots:**
   - Certain regions, like the **Middle East & North Africa** and **South Asia**, are hotspots for terrorist activities.

3. **Temporal Trends:**
   - The increasing trend in terrorist attacks up to **2014** and the sudden decline indicates the effects of global and regional policy changes and actions.

Limitations and Challenges:

1. **Data Completeness:**
   - Missing data, especially in 'Motive' and 'Summary', was a challenge, potentially skewing the analysis and leading to underestimation of certain dimensions of the dataset.

2. **Inherent Biases:**
   - The dataset primarily relies on publicly available information, which can introduce bias, especially in the context of politically sensitive topics such as terrorism. This might affect the accuracy and neutrality of the data.

4. **Correlation:**
   - The correlations are low. Therefore, the causes of trends or patterns cannot be conclusively determined without further in-depth study.

5. **Temporal Limitations:**
   - The dataset ends in 2017, and therefore, the findings may not fully reflect the current state of global terrorism dynamics, which are continuously evolving.