Report on Global Terrorism Database (GTD) Analysis

1. Introduction

The Global Terrorism Database (GTD) is an extensive dataset that contains information on terrorist events worldwide from 1970 to the present. This project aims to analyze the GTD to uncover trends, patterns, and insights related to global terrorism. The analysis includes data acquisition and preprocessing, statistical analysis, visualization, and performance comparison between Pandas and Dask.

2. Data Acquisition and Preprocessing

- **Dataset Loading**: The GTD dataset was loaded into a Pandas DataFrame for analysis.
- **Exploration**: Initial exploration revealed the structure and features of the dataset, including columns such as iyear, region_txt, attacktype1_txt, and nkill.
- **Data Cleaning**: Missing values were handled appropriately. For instance, columns with significant missing data were dropped and important columns with less missing values, just missing rows dropped.

3. Data Analysis

Statistical Analysis

- **Mean, Median, and Standard Deviation**: Calculated for numeric columns like nkill (number of kills).
  - Mean casualties: 2.12
  - Median casualties: 1.0
  - Standard deviation of casualties: 7.51
- **Most Frequent Values**: Identified the most common values in categorical columns.
  - Most frequent attack type: Bombing/Explosion
  - Most affected region: Middle East & North Africa
  - Most affected country: Iraq
  - Most common target: Private Citizens & Property

Grouping and Aggregation

- **Yearly Trends**: The number of terrorist attacks was grouped by year, revealing trends over time.
  - Significant increases or decreases in specific years were identified.
- **Regional and Country Analysis**: Grouped data by region and country to determine the most affected areas.

  - Most affected region: Middle East & North Africa (18.7K)

- **Attack Types and Targets**: Grouped by attack type and target type to understand common methods and targets of terrorist activities.
  - Most common attack type: Bombing/Explosion (55.6%)
  - Most common target type Private Citizens & Property

4. Data Visualization

Matplotlib and Seaborn

- **Trend Over Years**: A line plot was created showing the trend of terrorist attacks over the years.
  - Visualization revealed periods of increased terrorist activity.
  - Noticed that Trend of Terrorist Attacks increasing Over the Years
- **Attacks by Region and Country**: Bar plots were created to show the number of attacks by region and country.
  - Identified regions and countries with the highest number of attacks.
  - Noticed that Middle East & North Africa and South Asia are the highest numbers of attacks.
- **Correlation Heatmap**: A heatmap was generated to visualize the correlation between different features.
  - Identified significant correlations between variables.
  - Noticed that there are high correlations between numbers of kill and attack types.
- **Scatter Plot**: Showed the relationship between the number of casualties and the type of attack.
  - Insights into which attack types caused more casualties.

Plotly (Interactive Visualizations)

- **Geographic Distribution**: An interactive map showing the distribution of attacks globally.
  - Highlighted hotspots of terrorist activity.
- **Time Series Animation**: Animated the spread of terrorism over the years.
  - Visual representation of how terrorist activities have evolved over time.

5. Performance Comparison with Dask

- **Performance Measurement**: Compared the time taken by Pandas and Dask to perform groupby operations.
  1. Case #1: simple grouping by:

        - Pandas time: 0.0022 seconds
        - Dask time: 0.312 seconds

    2. Case 2#: Complex operation:
        - Pandas time: 5.82 seconds
        - Dask time: 2.75 seconds

- **Memory Usage**: Compared memory usage between Pandas and Dask.
    1. Dask showed better performance and memory efficiency with our datasets when use heavy operation.
    2. Dask uses lazy evaluation, meaning it doesn't read the file until you perform an operation that requires the data.

6. Insights and Findings

- **Trends Over Time**: Identified specific periods with significant increases in terrorist activities.
- **Regional Analysis**: Highlighted regions and countries most affected by terrorism, useful for policy and security measures.
- **Common Attack Methods**: Revealed the most common attack types and targets, providing insights into terrorist strategies.
- **Correlation Analysis**: Provided understanding of relationships between various features, aiding in deeper analysis and predictions.

7. Conclusion

The analysis of the Global Terrorism Database using Pandas, Dask, Matplotlib, Seaborn, and Plotly provided valuable insights into global terrorism trends and patterns. The use of Dask for large data handling demonstrated significant performance benefits but with our data pandas was good at simple grouping by and Dask was better when the operations becoming more complex. Visualizations enhanced the understanding of data, making it easier to interpret and communicate findings. This comprehensive analysis can inform policymaking, security measures, and further research in terrorism studies.