# Proposal: Analysis of Global Terrorist Attacks

**DATA 450 Capstone**

Shreehar Joshi

2/5/23

## 1 Introduction

Terrorist attacks - which are defined as the acts of violence committed by individuals or groups with the intention of causing harm and destruction to a large number of people - have been a major concern for countries across the world for many decades now. These attacks can take many shapes, including bombings, mass shootings, hijackings, and hostage situations, and can occur in any location, from urban centers to remote rural areas.

Terrorist attacks have a profound impact on both the victims and the larger society; they cause physical harm and loss of life, as well as emotional trauma and psychological distress. Needless to say, they can have long-lasting socio-economic consequences, disrupting trade and commerce, causing job losses, and decreasing investor confidence.

Many nations have increased their efforts to prevent and counteract terrorism in response to these attacks, which include stepping up their security protocols, improving their capacity for intelligence gathering and analysis, and cooperating with other nations to share information and coordinate efforts. Despite these initiatives, the threat of terrorism persists, with new organizations forming and established ones changing their strategies.

In this project, different terrorist incidents that occurred in different regions of the world from 1970 to 2017 will be analyzed for getting insight into various aspects of the incidents like the regions, methods, targets, motive, and the frequency of the terrorist attacks. In addition, the impact of these attacks on socioeconomic factors like the GDP, migration, and population will also be analyzed. Lastly, this project will assess the efficiency of different machine learning models in predicting the number of casualties from the existing incident data.

## 2 Datasets

In this project, three different datasets will be used to analyze different trends in terrorist activities and their impact on the social and economic health of different countries.

The first dataset - Global Terrorism Database (GTD) - is an open-source database tha includes information on more than 180,000 terrorist attacks around the world from 1970 through 2017. The database is maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland. Prepared by reviewing more than 4,000,000 news articles and 25,000 news sources, it is the most comprehensive unclassified database on terrorist attacks in the world. The variables that will be used in the analysis from this dataset are as follows:

- eventid: The 12-digit event id of the incident.
- iyear: The year in which the incident occurred
- country: The country code in which the incident occurred
- country_txt: The country name in which the incident occurred
- region_txt: The geographic region in which the incident occurred
- prov_state: The name of the 1st order subnational administrative region
- city: The city in which the incident occurred
- latitude: The latitude of the city in which the incident occurred
- longitude: The longitude of the city in which the incident occurred
- suicide: A flag indicating if the incident was a suicide attack
- attacktype1_txt: The general method of attack and broad class of tactic used
- targtype1_txt: The general type of the target/victim
- gname: The name of the group that carried out the attack
- motive: The motive of the terrorist attacks
- weaptype1_txt: The general type of the weapon used in the incident
- nkill: The total number of confirmed fatalities for the incident

The second dataset - World, Region, Country GDP/GDP per capita - is also an open-source dataset in kaggle consisting of the Gross Domestic Product data of different countries in the world from 1960 to 2021. It was created by Todor Mishinev using the World Bank National Accounts data. The variables that will be used from this dataset are as follows:

- Country Name: The name of the country
- Country Code: The three-letter codes for the countries
- year: The year for which the GDP data is recorded
- GDP_per_capita: GDP per capita for a specific country in the given year

Similarly, the third dataset - World Population - is an open-source dataset in kaggle consisting of different information on the world population from 1955 to 2020. It was created by Nguen Thi Cam Lai using the Countries in the world by population dataset from the Worldometer website. The variables that will be used from this dataset are as follows:

- Year: The year for which the population data is recorded
- Migrants(net): The net number of migrants in a country
- FertlityRate: The fertility rate of a specific country

# 3 Data Acquisition and Processing

For visualization, firstly, the variable names of all three datasets will be converted to camelcase to ensure uniformity. All the columns will be checked for the presence of null values and imputed with mean, median, and mode or their proportions (if applicable) with respect to the data types, the nature of distribution, and applicable subgroups. Any column, except the motive column from the first dataset, will be dropped if the percentage of null values exceeds 50. I will be spending a significant amount of time standardizing the names of the countries in each of the datasets; dataset 1 will be taken as the reference. This will not only ensure that the process of joining datasets is easier but also avoid any confusion or conflicts regarding countries' names in the visualization of data from different datasets.

For the modeling phase, all the categorical variables like target types, names of the terrorist groups, countries, regions, etc will be encoded to numbers and the motive variable will be processed through Term Frequency - Inverse Document Frequency vectorization to extract the most common words. The processed data will then be split and used to train and test the models.

# 4 Research Questions and Methodology

The following questions will be answered in this project:

### 4.0.1 Focus on the US

1. Which states in the US had the highest number of terrorist attacks? To answer this, I will first, create a dataframe - using the groupby and count function based on the provstate and event id respectively - that lists the number of terrorist attacks in each state, filter the dataframe to include data for the US, and then plot a choropleth map that shows the frequency of occurence of attacks in each state. The gradient of the color for each state will be proportional to the number of attacks.
2. How has the motive of terrorist attacks shifted over the last five years in the US? To answer this, I will first, create a dataframe that consists of rows representing terrorist incidents in the US from 2012 - 2017 and then use the motive variable to plot the Word Cloud. The plot will have the most repeated words for the motive and the size of the

words will represent the frequency of their occurrence. There will be five Word Clouds altogether, representing different years and they will be in a facet grid.

#### 4.0.2 Focus on the world

3. How has the frequency of terrorist attacks changed over the last five decades? To answer this, I will first, create a dataframe - using the groupby and count function based on the year and event id respectively - that lists the number of terrorist attacks in each year and then plot an interactive line plot that shows the trend of the occurence. There will be a single line, representing the number of terrorist attacks for the years 1970 - 2017.
4. Which geographical region has had the highest number of terrorist attacks? To answer this, I will first, create a dataframe - using the groupby and count function based on region and event id respectively - that lists the count of attacks for each of the geographical regions and then plot a mosaic plot displaying the data. The plot will be filled with distinct colors for each of the region and the size of the boxes representing each region will be proportional to their number of attacks.
5. How does the frequency of attack types vary across geographical regions? To answer this, I will first, create a dataframe - using the groupby and count function based on region and attack type and event id respectively - that lists the count of different types of attacks for each of the geographical regions and then plot a stacked bar plot displaying the data. All the bars will have the same height and width and each of them will represent a distinct geographical region. The bars will be filled with sub-bar plots of different colors, each of which will represent the type of attack for the given region. The height of each of the sub-bar plots will be proportional to the frequency of the attack type.
6. Which countries had the highest number of terrorist attacks? To answer this question, I will first, create a dataframe - using the groupby and count function based on country and event id respectively - that lists the count of attacks for each of the countries and then plot a bar plot displaying the data. The plot will be filled with distinct colors for each of the top 10 countries in descending order of the counts of attacks from left to right.
7. Do the countries with the highest number of terrorist attacks also have the highest number of people who were killed? To answer this question, I will first, create a dataframe - using the groupby and sum function based on country and number of casualties respectively - that lists the number of people who were killed from each of the countries and then plot a bar plot displaying the data. The plot will be filled with distinct colors for each of the top 10 countries in descending order of the number of people who were killed.
8. What are the most common groups targeted by terrorists? To answer this, I will first, create a dataframe - using the groupby and count function based on the type of target and event id respectively - that lists the count of different types of targets and then plot

a pie chart displaying the data. The pie chart will have five of the most common target types, each filled with a distinct color, and the rest of the types will be represented as "Others".

9. Which terrorist group is the deadliest (in terms of the number of people killed)? To answer this, I will first, create a dataframe - using the groupby and sum function based on group name and number of people killed respectively - that lists the number of people killed by each terrorist groups and then plot a mosaic plot displaying the data. The plot will be filled with the same color but of different sizes for each terrorist group, where the size of the boxes will be proportional to the number of people they have killed.

#### 4.0.3 Focus on socioeconomic factors

10. Is there any correlation between the GDP per capita and the number of terrorist attacks in a country? To answer this question, I will create a dataframe (referred to as dfCount henceforth) by selecting five countries with the highest number of terrorist attacks and calculate the number of terrorist attacks for each year for these countries by using the groupby and count function in country and year and event id respectively. I will then join this dataframe with our second dataset based on year and country column. The resulting data will be used in a line plot where the line will represent the number of terrorist attacks for each year in a country. On the same line plot, another line indicating the GDP per capita of the country for each year will be plotted. The two lines will be of different colors and to make sure that the lines are not of irregular scale (which is highly likely as GDP and the number of terrorist attacks differ largely in their magnitudes), both of them will be standardized to represent the minimum value of each category as 0 and the maximum value as 1. Lastly, each of the countries will have its own line plot and altogether, there will be 5 line plots in a facet grid. The trends in these five terrorist-prone countries will give us a general indication of the existence of the relationship between GDP per capita and the number of terrorist attacks.

11. Is there any correlation between the net number of migrants and the number of terrorist attacks in a country? To answer this question, I will join the dfCount dataframe from 10 with our third dataset based on the year and country column. The resulting data (referred to as dfJoinedThird henceforth) will be used in a line plot where the line will represent the number of terrorist attacks for each year in a country. On the same line plot, another line indicating the net number of migrants in the country for each year will be plotted. The two lines will be of different color and to make sure that the lines are not of irregular scale (which is highly likely as the net number of migrants and the number of terrorist attacks differ largely in their magnitudes), both of them will be standardized to represent the minimum value of each category as 0 and the maximum value as 1. Lastly, each of the countries will have its own line plot and altogether, there will be 5 line plots on a facet grid. The trends in these five terrorist-prone countries will give us a general

indication of the existence of the relationship between net number of migrants and the number of terrorist attacks.

12. Is there any correlation between the fertility rate and the number of terrorist attacks in a country? To answer this question, I will use the dataframe dfJoinedThird from 11 to plot in a line plot where the line will represent the number of terrorist attacks for each year in a country. On the same line plot, another line indicating the fertility rate of the country for each year will be plotted. The two lines will be of different colors and to make sure that the lines are not of irregular scale (which is highly likely as fertility rate and the number of terrorist attacks differ largely in their magnitudes), both of them will be standardized to represent the minimum value of each category as 0 and the maximum value as 1. Lastly, each of the countries will have its own line plot and altogether, there will be 5 line plots on a facet grid. The trends in these five terrorist-prone countries will give us a general indication of the existence of the relationship between the fertility rate and the number of terrorist attacks.

#### 4.0.4 Modeling

13. Can we predict the number of casualties from the existing features in the database? To answer this question, the variables of the first dataset listed in dataset section will be first checked for their "importance" using the feature importance feature of Random Forest Classifier. The top 10 variables in terms of importance will then be selected and label encoded if they are categorical in nature, and ultimately used to perform a train-test split in the ratio 80:20. The train dataset will be used to train five different machine learning models (Linear Regression, Support Vector Machine, Random Forest, XG Boost, and Ada Boost Classifier). The models will be tested on the test set with mean squared error, mean absolute error, and coefficient of determination as their evaluation metrics. 3-fold cross validation using grid search will be used to tune the hyperparameters of the aforementioned models. The results will be plotted on a grouped bar plot where each group will have three bars representing different evaluation metrics for a model. There will be five groups altogether.

## 5 Work plan

**Week 4 (2/6 - 2/12):**

- Animation of chloropeth map for the introduction (2 hours)
- Data Tidying (3 hours)
- Question 1 (2 hours)

**Week 5 (2/13 - 2/19):**

- Question 2 (2 hours)
- Question 3 (1 hour)
- Question 4 (1 hour)
- Question 5 (1 hour)
- Question 6 (1 hour)

**Week 6 (2/20 - 2/26):**

- Question 7 (1 hour)
- Question 8 (1 hour)
- Question 9 (1 hour)
- Question 10 (2 hours)
- Question 11 (2 hours)

**Week 7 (2/27 - 3/5):**

- Question 12 (2 hours)
- Question 13 (6 hours)
- Presentation prep and practice (4 hours)

**Week 8 (3/6 - 3/12):**

- Question 13 (5.5 hours)
- Presentation peer review (1.5 hours)

**Week 9 (3/20 - 3/26):**

- Code Revisions (3 hours)
- Poster prep (4 hours)

**Week 10 (3/27 - 4/2):**

- Code Revisions (2.5 hours)
- Peer feedback (2.5 hours)
- Poster revisions (2 hours)

**Week 11 (4/3 - 4/9):**

- Code Revisions (6 hours)
- Poster revisions (1 hour).

**Week 12 (4/10 - 4/16):**

- Code and Poster Final Revisions (7 hours)

**Week 13 (4/17 - 4/23):**

- Draft blog post (7 hours).

**Week 14 (4/24 - 4/30):**

- Peer feedback (2.5 hours)
- Blog post revisions (2 hours)

**Week 15 (5/1 - 5/7):**

- Final presentation prep and practice.

**Final Exam Week (5/8):**

## 6 References


Countries in the world by population (2023). Worldometer. Retrieved February 5, 2023, from https://www.worldometers.info/world-population/population-by-country/

Information on more than 200,000 terrorist attacks. Global Terrorism Database. Retrieved February 5, 2023, from https://www.start.umd.edu/gtd/

Lai, N. T. C. (2023, February 3). Word population (1955-2020). Kaggle. Retrieved February 5, 2023, from https://www.kaggle.com/datasets/nguyenthicamlai/population-2022

Mishinev, T. (2022, September 9). World, region, country GDP/GDP per capita. Kaggle. Retrieved February 5, 2023, from https://www.kaggle.com/datasets/tmishinev/world-country-gdp-19602021

National Consortium for the Study of Terrorism and Responses to Terrorism. Global terrorism database. Kaggle. Retrieved February 5, 2023, from https://www.kaggle.com/datasets/START-UMD/gtd

World Bank. GDP (current US$). GDP National Accounts. Retrieved February 5, 2023, from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD