What are the possibilities of a person in their country becoming a victim of terrorism?

First Author, *Joseph Veltri (*jvelt1@unh.newhaven.edu) *Another Day, Another Data*
Second Author, *Khadija Khorakiwala (*kkhor1@unh.newhaven.edu) *Another Day, Another Data*
Third Author, *Akhila Pitla* (apitl1@unh.newhaven.edu) *Another Day, Another Data*

***Abstract* - This paper Is taking the dataset that we have selected and creating a pattern to see which countries are attacked from terrorism the most often. Also, to make a hypothesis for if there are certain reasons why these countries are attacked as often as they are. The database that we chose has a huge number of entries with a significant quantity of information that is missing. Even though the GTD contains several reports of attacks, the quality of this information may have an impact on subsequent extreme value analysis of extreme event. It is logical to conclude that terrorism will impact areas which are not independent as they are easier to manipulate and dictate over. Also, terrorists around the world prefer using firearms that are minimal but cause a large impact, which was proven in our experiments. When comparing the behavior of the major terrorist groups, it is apparent that they operate in areas with a low level of freedom; also, countries with a low level of religious freedom have more terrorist activities on their territory. Rival religious groups are frequently classified as terrorists, this can happen. The higher the freedom index, the less attacks occur in countries, terrorism shows a correlation with freedom. Based on the Global Terrorism Database, the bivariate analysis and decision tree classifier were used to evaluate the terrorist attacks from numerical and categorical aspects.**

I. Introduction

Our topic for this project is to take the dataset that we have selected and create a pattern to see which countries are attacked from terrorism the most often. Also, to make a hypothesis as to if there are certain reasons why these countries are attacked as often as they are. The data contains information from 1970-2017 which includes the type of terrorism event, the exact date it happened (month, day, and year), the exact location it happened (city and country), longitude and latitude, a summary of the event, the type of attack, location it was targeting (what kind of organization, business, etc), and who the attack was from.

II. Related Work

1. “The Transnational Dimension in Political Risk Analysis” by Rafaele Marchetti and Mattia Vitale published in 2013. The purpose of this paper is to introduce a new model of analysis that – taking into consideration current concepts of political risk and modern theories of globalization – integrates in a comprehensive framework the more traditional variables of political risk with a new transnational variable. In this article, the authors used the data from the Organization for Economic Co-Operation and Development (OECD) with metrics such as political stability, social tensions, expropriations, political violence, and transfer risk. The article also made charts and graphs with risk values with cause-and-effect approaches.

2. The global terrorism database - accomplishments and challenges published by Gary LaFree in March 2010 describes the original data collection efforts and the strategies implemented to improve the quality and comprehensiveness of the data. It also provides statistical analysis on the data. Firstly, the provide the top 20 most frequently attacked countries based on the terrorist activity published in newspapers and other media. It also shows different patterns of terrorism from 1970 to 2010. The paper concludes by providing 3 areas for future research: validation studies, expanding databases beyond completed terrorist attacks, and geospatial analysis.

3. The paper 'Quantitative Analysis of Global Terrorist Attacks Based on the Global Terrorism Database' written by authors Zhongbei Li and Xiangchun Li focus on top ten terrorist attacks with the highest degree of hazard in the past two decades by using K-means cluster analysis method. The terrorists were classified according to region, type of attack, type of target and type of weapon used by them. In conclusion research results help people get a clearer understanding of terrorism and improve the government’s vigilance and emergency management capabilities.

4. Machine Learning Techniques to Visualize and Predict Terrorist Attacks Worldwide using the Global Terrorism Database - Enrique Lee Huamaní, Alva Mantari Alicia, and Avid Roman-Gonzalez. Research focuses on one of the branches of artificial intelligence, which is Data Mining and Machine Learning. The idea is to use these techniques to visualize and predict possible terrorist attacks using classification models, the decision trees, and the Random Forest. The input would be a database that has a systematic record of worldwide terrorist attacks from 1970 to the last recorded year, which is 2018. As a result, it is needed to know the number of terrorist attacks in the world, the most frequent types of attacks, and the number of seizures caused by region; additionally, to be able to predict what kind of terrorist attack will occur and in which areas of the world. Finally, this research aims to help the scientific community use artificial intelligence to provide various types of solutions related to global events.

5. A Method for Classifying Events from the Global Terrorism Database. This article by Richard E. Berkibile seeks to address the problem by proposing a method for refining original Global Terrorism Database (GTD) data into a constructively valid, cross-national domestic terrorism dataset. The analysis begins with the definition of terrorism and further develops it by conceptually distinguishing its domestic and transnational forms. Because the GTD includes non-terrorist events and conflates transnational and domestic incidents, its raw form is unsuited for domestic research. Hence, the article examines common definitional attributes from terrorism and domestic terrorism literature. It concludes by specifying steps for assembling a dataset and examining descriptive statistics of the resulting population.

III. Proposed Method

The group’s proposed method is what are the possibilities of a person in their country becoming a victim of terrorism. The

plan that our group has is to solve our hypothesis by using the database, the literature review articles that we have found, and statistical analysis to create a hypothesis and answer our hypothesis. The group will analyze and classify each attack and access the risk and make more information/data based on that risk. Depending on the level of risk, it will determine what kind of terrorism is in that country.

Total number of incidents	Over 87000
Incident types include	38,000 bombings 13,000 assassinations 4,000 kidnappings
Minimum number of variables	45
Maximum number of variables	>120
Supervised by	12 Terrorism research experts
Sources of information	3,500,000 News articles, 2500 News sources

TABLE I: A BRIEF DESCRIPTION OF DATA

The database that we have selected contains data from the time of 1970- 2017 (except 1993) that includes the type of terrorism event, the exact date it happened (month, day, and year), the exact location it happened (city and country), longitude and latitude, a summary of the event, the type of attack, location it was targeting (what kind of organization, business, etc), and who the attack was from. The database is downloadable and is very accessible. The data that we have chosen is from the GTD, or Global Terrorism Database and it is produced and maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START). This includes the data and codebook, any auxiliary materials present, and the user interface by which the data are presented.

	LOCATION					**ATTACK_TYPE**					**TARGET**				
	EUROPE	ME & NA	ASIA	AMERICA	RUSSIA	ASSAS	ARM_ASS	BOMBING	HOSTAGE	FACILITY	CIVILIANS	GOV	MILITARY	BUSINESS	OTHER
Occ (M)	227	**265**	115	132	9	36	**230**	270	43	148	199	237	**157**	97	59
%	30	**36**	15	18	1	5	**32**	37	6	20	26	32	**21**	13	8
Total	748					748					748				
Nb_Kills	2668														
Nb_Wounds	5599														

TABLE II. NUMBER OF OCCURRENCE OF MODALITIES IN DATASET

IV. METHODOLOGY

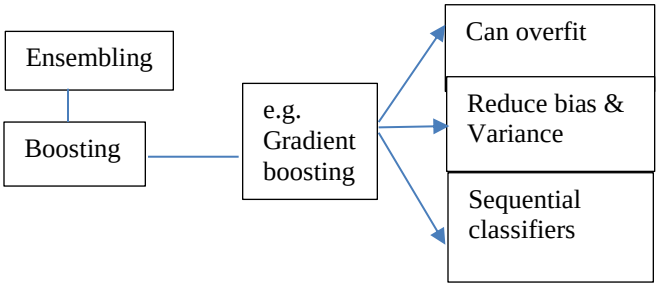
A. Data Cleaning/Preprocessing:

The database that we chose has a huge number of entries with a significant quantity of information that is missing. Despite the fact that the GTD contains several reports of attacks, the quality of this information may have an impact on subsequent extreme value analysis of extreme event. There are many missing records on one or more explanatory factors, such as the number of fatalities and the time stamps of events, for example. Additionally, many of the spatial locations of events are erroneous, with many entries' specific positions being either inaccurate or approximated using the longitudes and latitudes of nearby cities. Furthermore, it is clear that many countries, particularly those that experienced assaults prior to 2000, underreported the data in terms of severity and reported causalities. This would obviously pose issues with any analysis done during these time periods. To obtain reliable results, we applied a variety of preprocessing approaches to deal with missing data and inaccurate values in the dataset.

B. Gradient Boosting Algorithm:

One of the most powerful algorithms in the field of machine learning is the gradient boosting algorithm. As we know, machine learning algorithm errors can be divided into two categories: bias error and variance error. Gradient boosting is one of the boosting algorithms that would be used to reduce the model's bias error.

Gradient boosting recasts boosting as a numerical optimization issue in which the goal is to reduce the model's loss function by using gradient descent to add weak learners. Gradient descent is a first-order iterative optimization algorithm for determining a differentiable function's local minimum. Gradient boosting is a flexible technique that can be applied to regression, multi-class classification, as well as other applications because it is based on minimizing a loss function.



Gradient boosting, like other boosting approaches, iteratively merges weak "learners" into a single strong learner. It's easiest to understand in the context of least-squares regression, where the goal is to "teach" a model $\boldsymbol{F}$ to predict values of the form by $\boldsymbol{\bar{y}=F(x)}$ minimizing the mean squared error , $\frac{1}{n}\sum_i \widehat{(y_i} - y_i)^2$ where $\boldsymbol{n}$ indexes over some training set of size $\boldsymbol{n}$ actual values of the output variable $\boldsymbol{y}$:

- $\hat{y}_i$= The predicted value of $\boldsymbol{F(x)}$.
- y_i= The observed value.
- n = The number of samples in $\boldsymbol{y}$.

Now, let us consider a gradient boosting algorithm with $\boldsymbol{M}$ stages. At each stage $\boldsymbol{m}$ ($\boldsymbol{1<m<M}$) of gradient boosting, suppose some imperfect model $\boldsymbol{F_m}$ (for low $\boldsymbol{m}$, this model may simply return $\hat{y}_i = \bar{y}$, where the RHS is the mean of $\boldsymbol{y}$). To improve $\boldsymbol{F_m}$ our algorithm should add some new estimator, $\boldsymbol{h_m(x)}$. Thus,

$$F_{m+1}(x)=F_m(x)+h_m(x)=y$$

or equivalently,

$$h_m(x)=y-F_m(x)$$

Therefore, gradient boosting will fit $\boldsymbol{h}$ to the residual $\boldsymbol{y-F_m(x)}$. As in other boosting variants, each $\boldsymbol{F_{m+1}}$ attempts to correct the errors of its predecessor $\boldsymbol{F_m}$. A generalization of this idea to loss functions other than squared error, and to classification and ranking problems, follows from the observation that residuals $\boldsymbol{h_m(x)}$ for a given model are proportional equivalent to the negative gradients of the mean squared error (MSE) loss function (with respect to $\boldsymbol{F(x)}$):

$$L_{MsE} = \frac{1}{n}(y - F(x))^2$$

$$-\frac{\partial L_{MSE}}{\partial F}=2(y-F(x)) =\frac{2}{n}h_m(x)$$

So, gradient boosting could be specialized to a gradient descent algorithm, and generalizing it entails "plugging in" a different loss and its gradient.

Algorithm:

1.Initialize model with a constant value:

$$F_0(x)=arg\gamma\ min \sum_{i=1}^{n} L(y_i , \gamma)$$

2.For m = 1 to M:

A. Compute so-called pseudo-residuals:

$$r_{im}=-\left[\frac{\partial L(yi,F(x))}{\partial F(xi)}\right]_{F(x)-F(m-1)\,(x)} \quad for\ i=1,2,\ldots\ldots\ldots n$$

B. Fit a base learner (or weak learner, e.g., tree) closed under scaling $h_m(x)$ to pseudocode -residuals, i.e., train it using the training set $\{(x_i , \gamma_{im})\}^n_{i=1}$

C. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m=arg\gamma\ min \sum_{i=1}^{n} L(y_i , F_{m-1}(x_i)+\gamma h_m(x_i)).$$

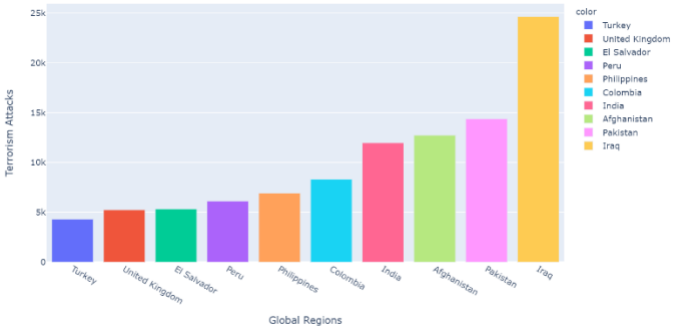
$$F_m(x)=F_{m-1}(x)+\gamma_m h_m(x_i)$$

3. **Output**: $F_M(x)$.

V. Experimental Results

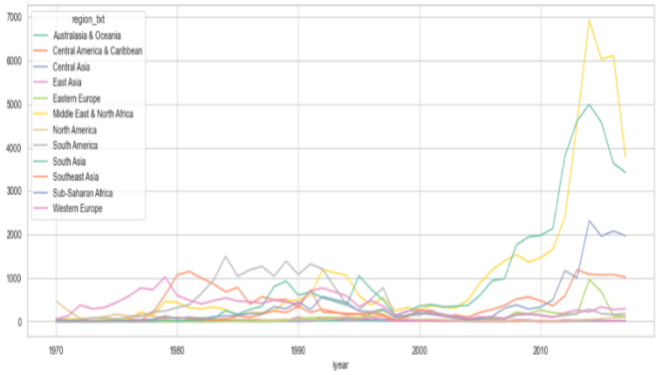
The experiments are carried on using GTD terrorism incident records from the years 2001 to 20017. Each terrorist record includes a news report as well as several other characteristics of terrorism incident including the type of incident.

A. The trend in the number of terrorist attacks by region:



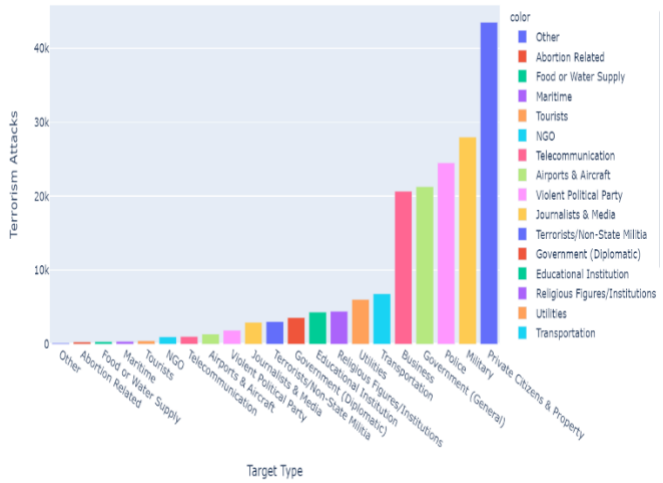
When looking at the first graph, it's notable to observe that there was a general low in terrorism from 1990 to roughly 2000, followed by a massive spike. This could have something to do with 9/11 and the US's entire war on terrorism.

B. How many people died relatively per country?



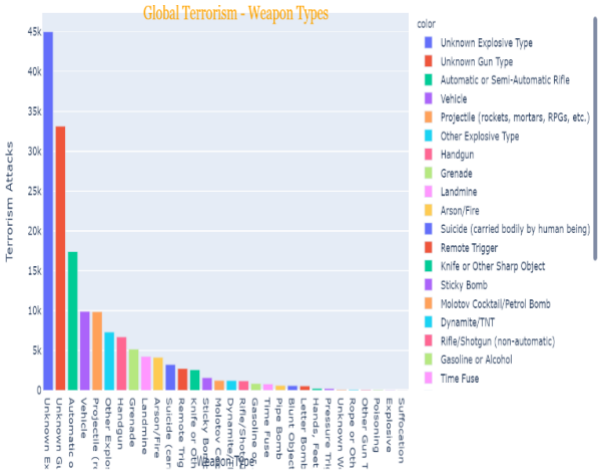
It's significant to note that the number of attacks and kills relative to the population is for the bulk of the countries almost the same. Some countries, however, have much greater numbers. The Falkland Islands have a relatively high number of documented attacks, but no known kills, implying that either no deaths have occurred, or no one has been killed during terrorist attacks. Iraq is the most hazardous country in the world, with the highest number of attacks and deaths per capita. Nicaragua and El Salvador are in second and third position, respectively. The most dangerous countries in Europe are the United Kingdom, Croatia, and Cyprus.

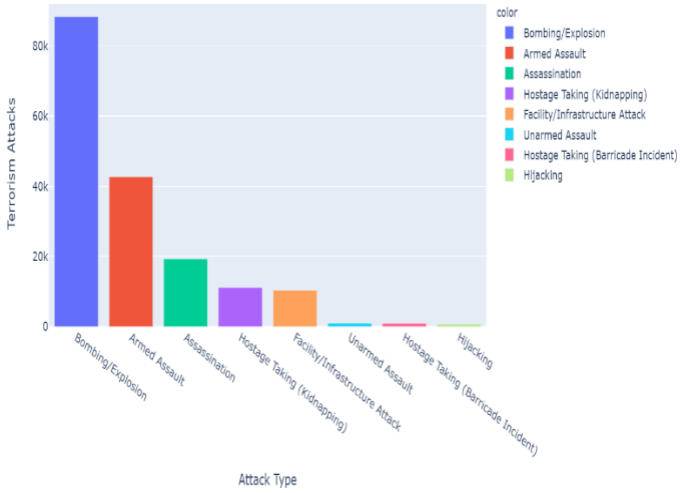
C. Terrorism impact on the target:



Private citizens, military, and police collectively account for approximately three-quarters of all kills. Property damage, on the other hand, is more uniformly spread among the various target types. Military attacks have the largest average number of deaths each attack, although the number of wounded every attack is relatively low when compared to the other target types. Terrorist attacks kill a greater number of defense personnel (police and military) than civilians. Property damage to defense groups is higher than that of private persons. When comparing the public and private sectors, the public sector outperforms the private sector. The public sector bears the brunt of the damage more than the private sector.

D. Terrorist acts have the greatest influence on which types of attacks?





Bombing has the highest number of wounded and overall property damage; the most significant difference is between hijacking and unarmed assault. When compared to bombing, an armed attack results in the same number of deaths, but less injuries and property damage. Infrastructure assaults cause a lot of property damage but just a few deaths and injuries.

VI. Discussion

The final results were cohesive with our expected results. It is logical to conclude that terrorism will impact areas which are not independent as they are easier to manipulate and dictate over. Also, terrorists around the world prefer using firearms that are minimal but cause a large impact, which was proven in our experiments. The top 3 areas which experienced major terrorism is Iraq, Pakistan, and Afghanistan. Once the resources of these areas have been exhausted, the hotspots can shift to other areas like India, Columbia, and Philippines (which can be seen in the graph). Therefore, the graph can be interpreted to predict the next likely hotspots for terrorism. The results showed that number of people injured is high in case of bombing and explosives. This information can be used to monitor the firearms in different countries, thereby helping to prevent future attacks.

VII. Conclusion/Future Work

When comparing the behavior of the major terrorist groups, it is apparent that they operate in areas with a low level of freedom; also, countries with a low level of religious freedom have more terrorist activities on their territory. Rival religious groups are frequently classified as terrorists, this can happen. The higher the freedom index, the less attacks occur in countries, terrorism shows a correlation with freedom. Based on the Global Terrorism Database, the bivariate analysis and decision tree classifier were used to evaluate the terrorist attacks from numerical and categorical aspects. The primary graph shows the terrorist activities by region in each year from 1970 to 2019. The number of attacks range from 0 to 16,000. By this we can conclude that the regions with the high number of attacks pose a higher risk level.

There are only a few countries that account for most of the

global terrorism attacks, and within these countries, only a few states or cities can be called hotspots for practically all attacks. Iraq and Pakistan are two of the world's most dangerous terrorist hotspots. Most Iraqi and Pakistani states, such as Baghdad in Iraq and Balochistan in Pakistan, were found in the top 10 global terrorism state-by-state study. In Pakistan and Iraq, groups like the Taliban and ISIL are the main perpetrators. According to the data, Iraq appeared five times in the top 20 terrorism attacks in terms of the number of people killed in each attack. Attackers mostly target 'Business,' 'Government (General),' 'Police,' 'Military,' and 'Private Citizens & Property,' among other things. The major attack types in over half of all terrorist cases are bombings or explosions. 'Unknown Explosive Type,' 'Unknown Gun Type,' and 'Automatic or Semi-Automatic Rifle' were the most common weapons they used. Some future work that the group would like to accomplish are expanding this project by including big data techniques to conduct sentiment analysis, thus extracting information from social networks that are a part of cyber terrorism.

VIII. GitHub Repository Link

https://github.com/joeveltri/Phase-8-FinalReport

IX. Resources

- R. Marchetti and M. Vitale, “The Transnational Dimension in Political Risk Analysis,” 2013.
- Z. Li, X. Li, C. Dong, F. Guo, F. Zhang, and Q. Zhang, “Quantitative Analysis of Global Terrorist Attacks Based on the Global Terrorism Database,” Sustainability, vol. 13, no. 14, p. 7598, 2021
- E. L. Huamaní, A. Mantari, and A. Roman-Gonzalez, “Machine Learning Techniques to Visualize and Predict Terrorist Attacks Worldwide using the Global Terrorism Database,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, 2020
- R. E. Berkebile, “What Is Domestic Terrorism? A Method for Classifying Events From the Global Terrorism Database,” Terrorism and Political Violence, vol. 29, no. 1, pp. 1–26, 2015.
- “Gradient boosting,” Wikipedia, 17-Nov-2021. https://en.wikipedia.org/wiki/Gradient_boosting. 06-Dec-2021

X. Proofreading from Writing Center