AIT622 - Determining Needs for Complex Big Data Systems

# GLOBAL TERRORISM ANALYSIS

## BIG DATA PROJECT MANAGEMENT PLAN

By,
Sai Hemanth Nirujogi,
G01065588

## Introduction

“What is terrorism?” is one of the most significant questions in today’s world. Terrorism itself is one of the most controversial and contested of the social sciences because it is so politically loaded. One of the drivers of terrorism are people’s reactions to military conflicts, whether it is occupation, or be an invasion of a particular country etc. Terrorism is not just related to one particular region or religion. And men and women are equally capable of conducting violence in the name of a terrorist or extremist cause. In 2015, there are almost 15,000 terror attacks in the world [1]. The problem is there is no internationally recognized term for terrorism. Terrorism has become a general term for various acts of violence, which are not even considered as terrorism before. The FBI defined terrorism as “the unlawful use of force or violence against a person or property to intimidate or coerce a government, the civilian population, any segment thereof, in furtherance of a political or social objective”. [2] In some cases, it may be obvious to call some event as terrorism, for others it may be a lot more muddled.

This all comes down to a “motive”. Was there a political or ideological agenda behind the attack? When compared the 10 year period of terrorist incidents happened before and after the year 2003, there are 20,451 incidents took place between 1993-2003 and the years following the invasion from 2003-2013 the number more than doubled (49,678 incidents). So what happened In the year 2003 that made the terrorist incidents catastrophically high? In the year 2003, the United States invaded Iraq and some argue that the war happened because of the radical geopolitics [3].

Understanding and analyzing various acts of global terrorism as a big picture creates an intact awareness among those interested to contribute their interpretations to this problem. In this project, I plan to study various terror attacks that took place across the globe through the years 1970-2017. The few variables that need to be intensively studied are the success rate of the attack, the weapon used, the number of injured and killed, type of attack, targeted agencies, groups of people, property loss, etc. By using the power of statistical analysis, we will see what are the key factors that lead to the success of a terror attack.

## Objective

The main aim of this project is to study the global terrorism database to perform relevant summary statistics and visualizations. Analyzing the dataset to determine which variables have the most effect on the terrorist incidents. Finding the weights of the attributes to apply them on the data to find the importance of the attribute on the incidents. To apply the insights from the study to create a strategy, like what factors should be taken into account in defense of terror attacks and hate crimes.

## Project Outline

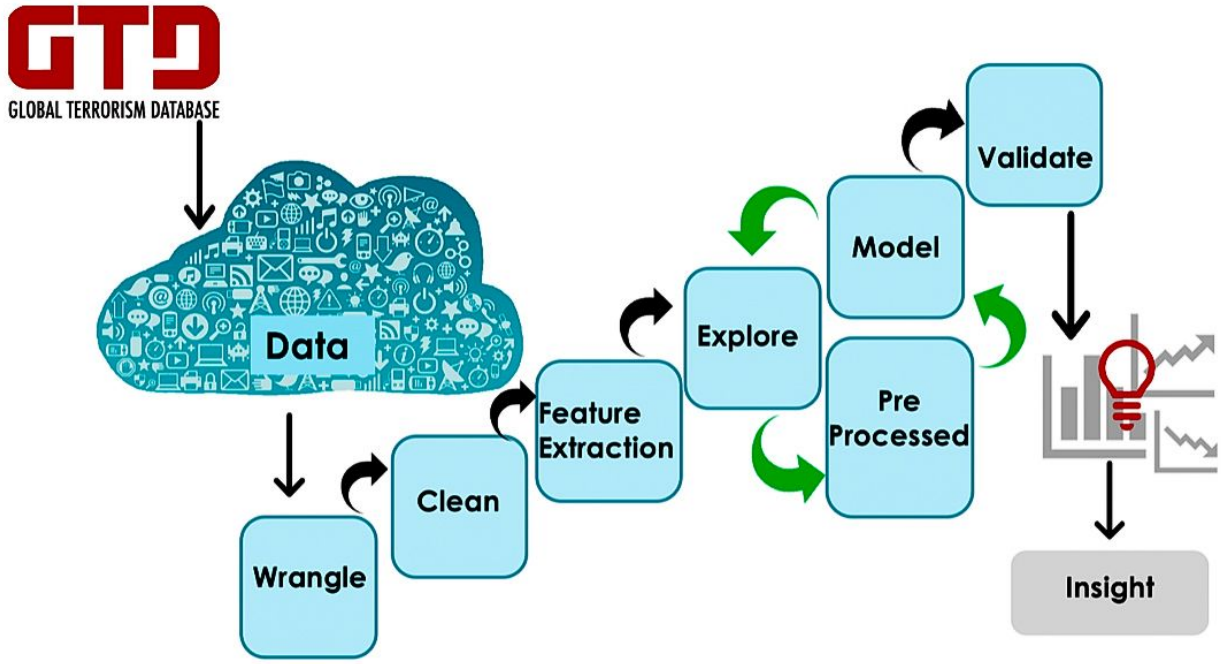



*Fig.1 Project Outline*

The project outline shows the steps we will be taking to find the insights from this data. The steps include data conditioning, feature extraction, data exploration, pre-processing, modeling, data validation and drawing conclusions from the results.

## Stakeholders

The data is collected by the Global Terrorism Database based at the National Consortium for the Study of Terrorism and Responses to Terrorism (START) located at the University of Maryland. START is comprised of a global network of merits committed to the study and human consequences of terrorism in the U.S and the rest of the world. They work with the Center for Terrorism and Intelligence Studies (CETIS) and funded by the Department of Homeland Security. Other organizations who helped in curating the data for the GTD include Institute for the Study of Violent Groups (ISVG) located at the University of New Haven. ISVG helped GTD in curated the data about the terrorist attacks happened between April 2008 and October 2011. In the Spring of 2012, GTD has moved all of its collection to the University of Maryland and the START team has made noteworthy improvements to the strategies they use to curate the database. The GTD is a collection of information from various sources and efforts since 1970 with data primarily built from the incidents documented in real-time by PGIS. [4]

## Data Quality and Privacy

The data is collected from various sources from all around the world by a lot of merit from different organizations. The data is collected entirely from public and open-source information made available to everyone. The sources include datasets curated by well-known sources, news archives, electronically available articles and secondary sources like journals, legal documents, and books. No matter how much effort is put into the data collection by well-trained merits, the data is still collected from open sources and independent journals and the veracity of the data is not completely claimed by those sources. The rights over the data are completely owned by START and no changes within the data should be made by any certain individual by inferring an individual associated with the incidents. New information and appropriate changes to the dataset will be only made available by the researchers from the START. There are 3 phases of data collection went in to make this database:

**Phase one (1970-1997):** Data is collected by Pinkerton Global Intelligence Service (PGIS), a private security agency.

**Phase two (1998-2008):** The incidents occurred between these years are identified and labeled by the Center for Terrorism and Intelligence Studies (CETIS), with association from START.

**Phase three (2008-2011):** Information between April 2008 and October 2011 is curated by ISVG, University of New Haven.

From the year 2011, START team has done the data collection for the GTD and worked on supplementing data on additional instances from various well-known archives till now. [5]

Data consistency is one of the issues START has been facing with GTD because of the collection of data between the years 1970-1997 is done in real-time as well as the period between 1998-2007 and again in 2007. The difference in the data collection is visible because the media sources throughout the years became unavailable. In the year 2012, START has created a methodology to compile the database. The model uses machine learning techniques to identify the articles related to terrorism and terror attacks. It collects all the information about the incident like location, time/date, number of people injured, cause etc. The GTD has no data of the year 1993 and still analyzing the insights provided by PGIS.

## Requirements

I. **People and HW/SW Resources**

- **A.** A lot of intellect and manpower has gone in curating the dataset from various sources around the world. Maintaining the data with same

consistency will also require the same amount of work to mine, condition and curate the data.

**B.** High-end hardware with high processing power, storage is required to maintain and compile such a complex database. Traditional tools like Python, R etc. Tableau and other tools for the visualization and statistical analysis are required.

**II. Build vs Buy Solutions**

Models and algorithms can be build simply by using traditional and powerful tools. To maintain the database in the cloud START could upgrade to new cloud solutions like AWS, Cloudera or Salesforce though it is not necessary. Even though the data is sometimes updated in real-time, it does not always require powerful cloud solutions.

## Metadata

The dataset consists of 180,000 rows and 89 self-explanatory variables with a significant amount of missing values. Details about each incident are included like date, time, lat and long data, target, the number of fatalities, motive, and type of weapon used in the attack. Each attack has at least 45 variables to describe it.

| Spatio-Temporal Variables | Continuous Variables | Binary Variables | Categorical Variables | Descriptive Variables |
|---|---|---|---|---|
| Iyear, imonth, iday, latitude, longitude | nkill, nwound | Crit1, crit2, crit3, doubtter, extended, multiple, success, suicide, guncertain1, claimed, property, ishostkid | Country_txt, region_txt, alternative_txt, attacktype1_txt, targtype1_txt, natlty1_txt, weaptype1_txt | target1, gname, summary |

The most important variables we should be focusing on are success, attach type, weapon type, target type, target subtype, latitude, longitude, nationality, number of attackers, number of casualties, attacker group, is host a kid?, us_casualities, region, corp name, motive, property loss, injured number etc.

The graph on the right represents the missing values in each column of the data set. As we can observe from the plot that weaptype, weapsubtype, claimode, attacktype3, targettype, guncertain, ransomaid and hours have the highest percentage of missing values.

This data needs to be imputed from the dataset to perform the analysis further. The variables with the high number of missing records are removed using a simple imputation technique. After the columns are removed from the dataset, we have to remove any null records.

Before removing the variables, we have to check for the near-zero variance because imputing the variables is not efficient because some of the variables are categorical. The near-zero variance predictors will be stored.

The imputation is done by dividing the dataset into numerical and charter set. The numerical data is assigned with 0 and the character variable is imputer with “MISSING”.

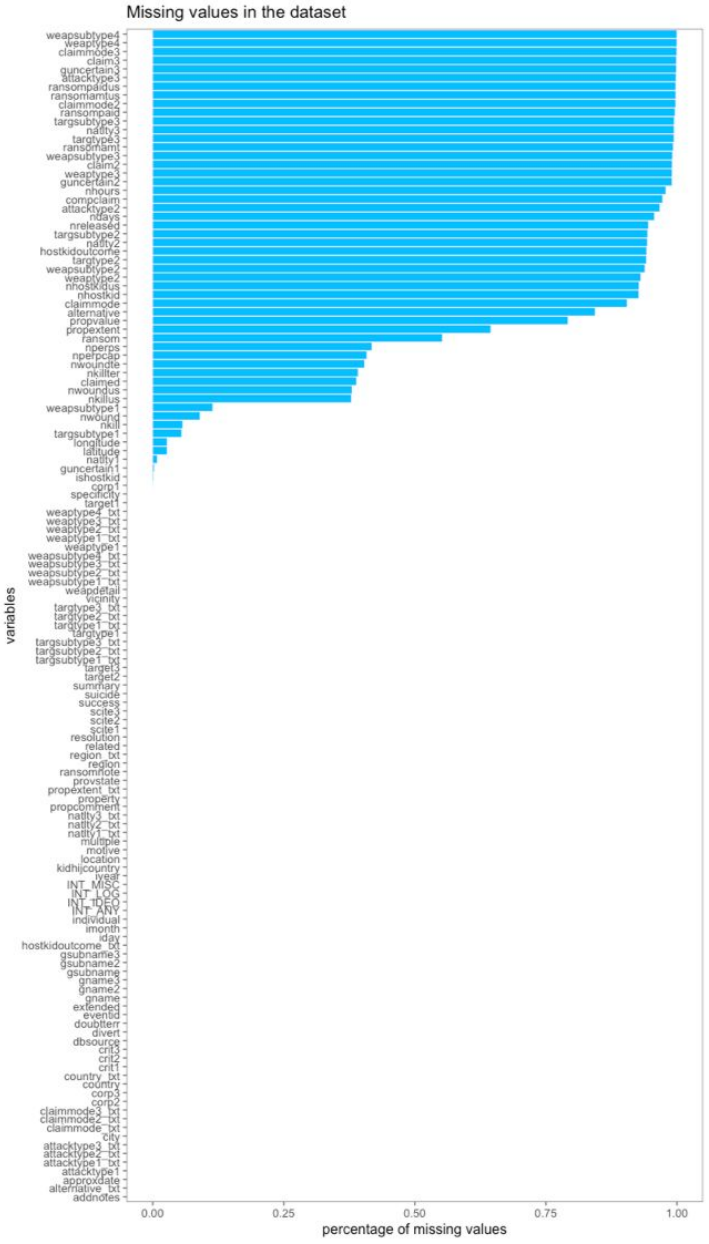



*Fig.2 Variables with missing values.*

```
for (x in attr.data.types$integer) {
  sample.df[[x]][is.na(sample.df[[x]])] <- -1
}

for (x in attr.data.types$character) {
  sample.df[[x]][is.na(sample.df[[x]])] <- "*MISSING*"
}
```

*Fig.3 Imputing numerical and categorical data*

## Statistical Analysis

We can perform various statistical analysis of this data to gain some insights. Various classification techniques and predictive analytics can be applied to this dataset.

### I. Feature Selection

Feature selection is done to find the variable importance and to find whether the variable can be accepted or rejected for the analysis. By using Boruta Analysis we found out that variables like weaptype4txt, claimmode2txt, weapsubtype4txt are rejected as shown in the snippet (fig.4) on the right.

```
> print(boruta.df)
                    meanImp   medianImp      minImp      maxImp normHits  decision
region           19.3365367 19.0948565 17.483995 21.2499410        1 Confirmed
regiontxt        19.4664906 19.4619002 18.350078 20.7252417        1 Confirmed
vicinity          5.7033960  5.9341725  4.698636  6.4811779        1 Confirmed
crit1             4.3934203  4.3461597  3.738290  5.1693824        1 Confirmed
crit2             6.3219666  6.9766145  3.513173  8.1463748        1 Confirmed
crit3             5.8479054  5.8032675  4.692426  7.5022363        1 Confirmed
doubtterr        11.0670993 11.4910713  6.547986 14.7500472        1 Confirmed
alternativetxt    6.7398491  6.8638889  4.323167  8.4615573        1 Confirmed
multiple         15.9718683 16.0431494 14.745613 16.9730491        1 Confirmed
success         235.2386742 235.1420589 224.856333 248.7956342        1 Confirmed
suicide           5.7429594  5.7275487  4.607971  6.6312405        1 Confirmed
attacktype1      21.5643312 21.6785962 19.827083 23.5762780        1 Confirmed
attacktype1txt   20.0404658 19.8332211 18.435616 21.5173115        1 Confirmed
attacktype2txt    8.1808093  7.8171177  7.418455 10.3131713        1 Confirmed
attacktype3txt    3.1330033  3.1905227  2.037196  3.9678736        1 Confirmed
targtype1        19.8912475 19.8045044 18.845396 22.2381104        1 Confirmed
targtype1txt     19.1660676 19.0658050 18.403336 20.0802048        1 Confirmed
targtype2txt     14.5665576 14.6954662 13.369854 15.3141869        1 Confirmed
targtype3txt      4.7936828  4.9160413  3.447014  5.7833737        1 Confirmed
guncertain1       9.5824344  9.5987110  8.174779 10.9891086        1 Confirmed
individual        4.8197416  4.7878829  2.789352  5.9358142        1 Confirmed
claimmodetxt     12.3336938 12.3944784 11.286508 13.2785653        1 Confirmed
claimmode2txt     0.2190405  0.1668344 -1.107344  1.5330791        0  Rejected
weaptype1        14.3066837 14.4779225 12.566673 15.4254172        1 Confirmed
weaptype1txt     16.8875622 16.7984788 14.098251 18.4624652        1 Confirmed
```

*Fig.4 Variable importance using Boruta Analysis*

### II. Random Forest

There are many types of classification models and amongst them Random Forest one of the most popular models. Random Forest is an ensemble classifier made using masing decision tree models. Using the Random Forest model we can find the variable importance. During the analysis, the Random Forest model displayed the best accuracy compared to other classification models. From the variable importance plot (fig.5) on the right, we can see that property, targtype1_txt, targtype1, weapsubtype1_txt, attacktype1 etc.

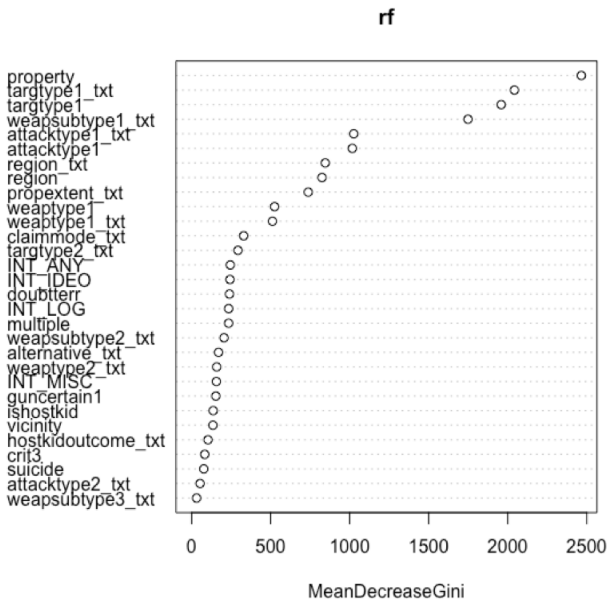



*Fig.5 Random Forest - Variable Importance*

## III. Classification Decision Tree

Another machine learning classification model includes Decision Tree. This type of classification models builds classification in the form of trees by breaking the dataset into smaller subsets. Decision tree uses a tree structure to specify decisions and consequences. Using tree-based classification on our data predicts which variable has the highest level of importance. We can see from fig. 6 that property has the highest level of importance.

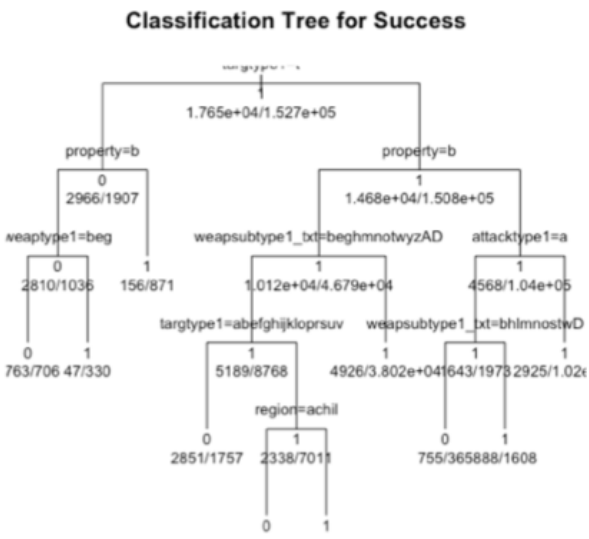



*Fig.6 Classification Decision Tree*

## IV. Logistic Regression

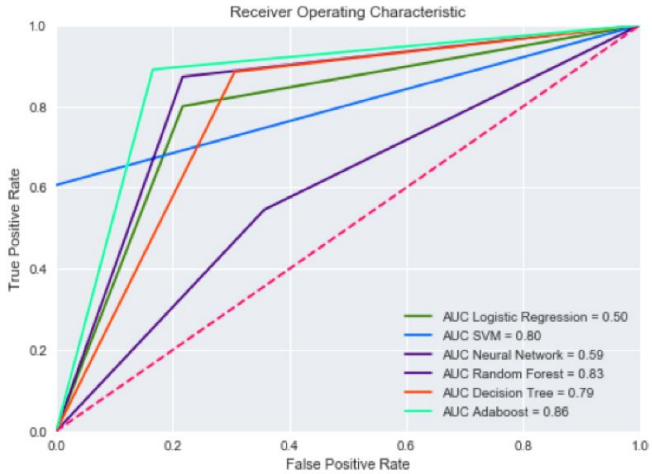
With a wide variety of categorical variables in our dataset Logistic regression can be the best analysis method we can apply.

Logistic Regression: 0.79%

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.79 | 0.78 | 0.79 | 157 |
| 1 | 0.80 | 0.80 | 0.80 | 165 |
| avg / total | 0.79 | 0.79 | 0.79 | 322 |

*Fig.7 Logistic Regression*

Six models are trained using the data, including, Support Vector Machine (SVM), Mult Layer Perceptron, AdaBoost and the summary of precision, recall, and accuracy are as follows:




| Model | precision | recall | accuracy |
|---|---|---|---|
| Logistic Regression | 79 | 79 | 79 |
| SVM | 86 | 80 | 80 |
| MLP | 59 | 58 | 58 |
| Random Forest | 82 | 81 | 81 |
| Decision Tree | 76 | 75 | 75 |
| Adaboost | 86 | 86 | 86 |

*Fig.8 ROC Curve and Summary of All the Models*

Other statistical tests include the Chi-Squared test, PCA, SVC etc. can be applied to the dataset to find the best variables that have the most effect on the success of the attack.

## Visualizations

### I. Correlation Plot

The plot shows the correlation between the numerical variables and there is very less correlation between the variables. The only high correlation is observed between the weapon type and suicide.

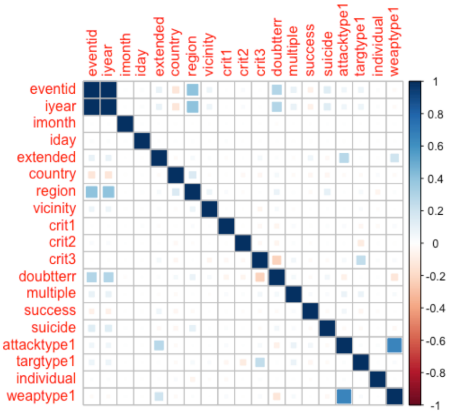



*Fig.9 Correlation Plot*

### II. Histogram

#### A. Weapon type used in the attacks

As we can clearly see from the graph explosives/bombs/dynamites are highly used in the attacks.

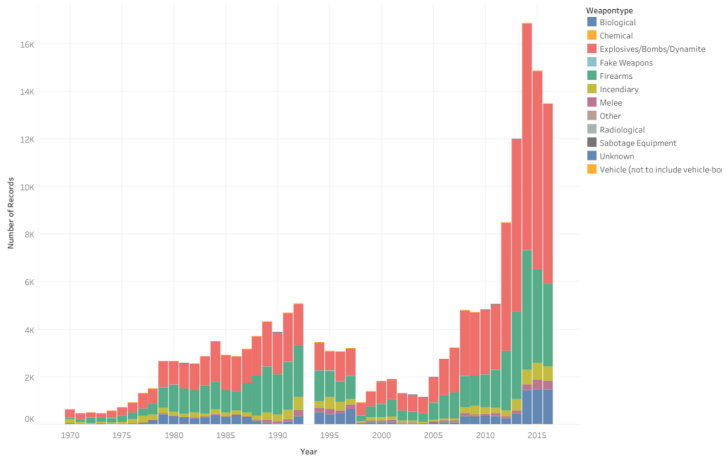



*Fig.9 Most used type of weapon in the attacks*

#### B. Number of fatalities in each country

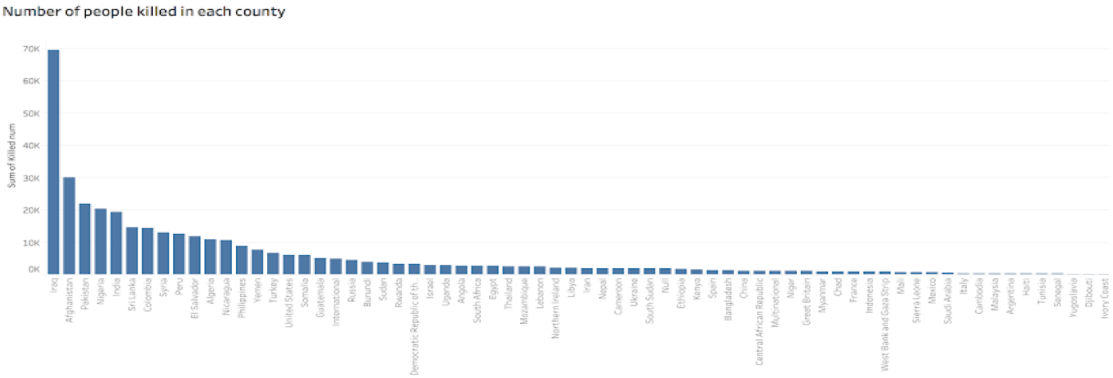



*Fig.10 Highest number of people killed in each country*

### C. Successful terror attacks in various countries/regions

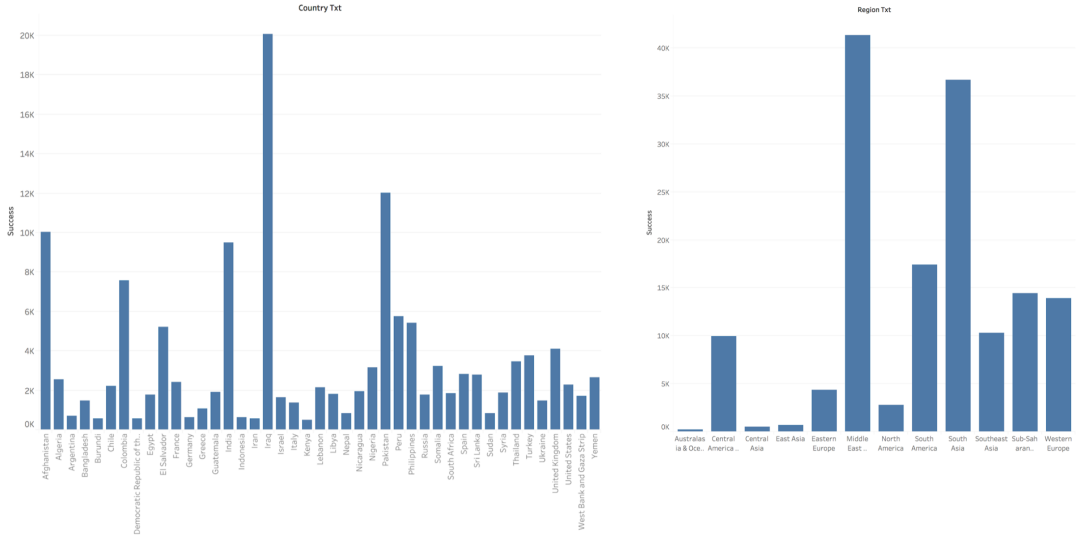



*Fig.11 Successful terror attacks in various countries/regions*

From the graph, we can see that the Middle East, South Asia, North Africa, and South America has the highest number of successful terror attacks.

## III. Visualization on a Map

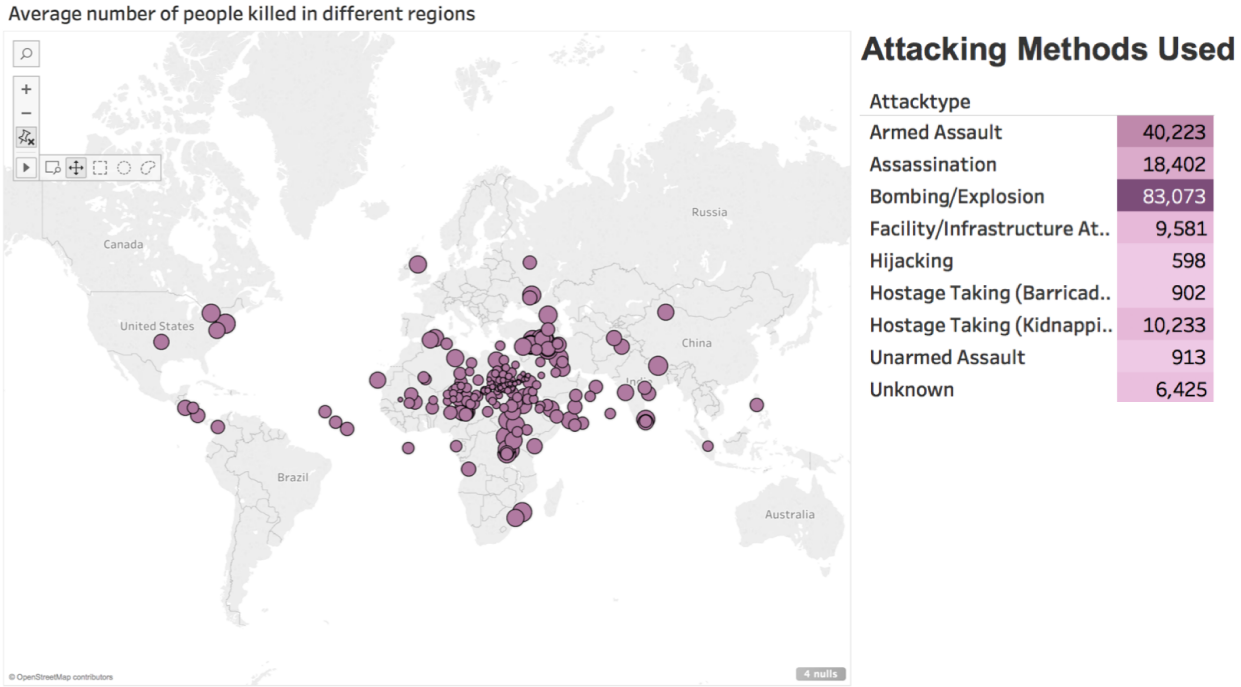



*Fig. 12 Average number of fatalities in each country by different attacking methods*

## Benefits/Expected Value

Analyzing a public dataset about a real-world issue can be helpful to the world in a lot of ways. The main purpose of this database collection in the first place is to predict the next attack based on the trends in the data and to stop it before happening. By analyzing this dataset we have found a few valuable insights

1. The variables like property, weapon type, target are the most important things to consider while determining the success of a terrorist attack.
2. Details of the region like latitude and longitude data, street, crowd etc., can be really helpful to stop an attack.
3. The variable incident showing the vicinity of the city also has high importance, as well as the attack type.

By analyzing the important variables, one can determine the success rate of an attack. The data and the insights can further put to future use by research organization to take necessary measures to reduce the consequences of an attack.

## Relevant Terms

I. GTD - Global Terrorism Database
II. START - Study of Terrorism and Responses to Terrorism
III. CETIS - Center for Terrorism and Intelligence Studies
IV. ISVG - Institute for the Study of Violent Groups
V. PGIS - Participatory Geographical Information Systems

## References


1. “Fewer People Are Dying from Terrorism Worldwide. But Right-Wing Terror Still Threatens the U.S.” U.S. News & World Report, U.S. News & World Report, accessed on 6th Dec. 2018 www.usnews.com/news/best-countries/articles/2018-12-05/global-terrorism-deaths-down-globally-right-wing-terror-on-rise.
2. Lazreg, Houssem Ben. “The Debate over What Constitutes Terrorism.” The Conversation, The Conversation, 27 Nov. 2018, accessed on 6th Dec. 2018 theconversation.com/the-debate-over-what-constitutes-terrorism-86812.
3. Mercille, Julien. “The Radical Geopolitics of US Foreign Policy: the 2003 Iraq War.” GeoJournal, vol. 75, no. 4, 2010, pp. 327–337. JSTOR, JSTOR, accessed on 6th Dec. 2018 www.jstor.org/stable/41148401.
4. “About GTD.” County-Level Correlates of Terrorist Attacks in the United States | START.umd.edu, accessed on 6th Dec. 2018 www.start.umd.edu/gtd/about/History.aspx.
5. “Using GTD.” Global Terrorism Database, accessed on 6th Dec. 2018 www.start-dev.umd.edu/gtd/using-gtd/.