**EDA on Global Terrorism (1970-2017)**

**Team Data titans**

**ABSTRACT**

An analysis of a project is performed, which consists of a variety of studies and visualisations for interpreting patterns and trends. The visualisation tool that gives the user the ability to explore datasets. Diverse viewpoints result from a lack of comprehension and awareness of global terrorism and widespread misunderstandings among civilians. In this age of globalisation, there are adequate knowledge on this topic which can help boost our anti-terrorism measures, security issues, enforce improved economic policies, and expand the body of information about civilians.

Alma Better has provided us with the primary dataset for this research, which is derived from a data source managed by START and contains information on terrorist attacks from 1970. Using a variety of data mining and data visualisation approaches to interpret the nature of terrorism in order to better comprehend its trends and patterns over its 45-year recorded history.

**PROBLEM STATEMENT**

We have been given a data set which has all the records of Global terrorism activities happening across the world from 1970-2017. We have performed several analysis to derive important information from the dataset.

**Data Dictionary:**

**Year:** The year in which the incident occurred.

**Month:** The number of the month in which the incident occurred.

**Day:** Day of the month on which the incident occurred.

**Country:** The country or location where the incident occurred.

**Region:** The region in which the incident occurred. The regions are divided into the 12 categories.

**City:** The name of the city, village, or town in which the incident occurred.

**Latitude:** the latitude of the city in which the event occurred

**Longitude:**the longitude of the city in which the event occurred

**Success:** it is defined according to the tangible effects of the attack

**Attack\_Type:** the general method of attack and often reflects the broad class of tactics used

**Target:** targets/victims is recorded for each incident

**Target\_Sub:** The target subtype variable captures the more specific target category and provides the next level of designation for each target type

**Terrorist\_Group:** the name of the group that carried out the attack

**Motive:** a specific motive for the attack

**Weopon\_Type:** the weapons used in an attack are recorded for each case

**State:** State of the country where the incident occurs

**Weopon\_Subtype:** subtypes of the weapons used in an attack are recorded for each case

**Killed:** the number of total confirmed fatalities for the incident

**Wounded:** the number of confirmed non-fatal injuries to both perpetrators and victims

**Property:** if there is evidence of property damage from the incident

**Property\_Damage:** the following four categories describes the extent of the property damage: 1 = Catastrophic (likely = 1billion)2=Major(likely=1 million but < 1billion)3=Minor(likely<1 million) 4 = Unknown

**Database\_Source:** the original data collection effort in which each event was recorded

**Addnotes:** additional relevant details about the attack

**Casualty:** totality of killed and wounded people

**Counts:** Number of deadliest attacks

**INTRODUCTION**

World peace was one of the primary motivations for establishing the United Nations [1]. Terrorism is the greatest obstacle to global peace. Terrorism is frequently disregarded by those who are not immediately impacted by its dangers. Terrorism is generally regarded as an unpredictable and tragic tragedy that disproportionately affects certain regions of the world. On the basis of the location of the incidents, the general public has extremely limited information about similar occurrences in other areas of the world and therefore responds differently [2]. This research focuses on terrorism by analysing the dataset supplied by an unidentified source in order to discover significant trends and statistics. It's difficult to define terrorism. The United Nations General Assembly is unable to come to an understanding on a single definition of terrorism at this time [4]. Different governments and organisations define terrorism differently as a result of this issue. This ambiguity generates many conflicts regarding which events constitute terrorism and which do not. Various organisations define terrorism differently and conduct their business appropriately. As a result, the contents of terrorism-related datasets collected by independent groups may differ significantly [5]. Consequently, the analyses and outcomes of this study may differ from identical studies performed on a different dataset.

**MISCONCEPTIONS ABOUT TERRORISM**

Terrorism is intermittent, pervasive, and inconsistent with time and nature. Due to these qualities, it is impossible to synthesise all aspects of international terrorism into a single, convincing solution and make this knowledge accessible to the general public [6]. Exploring this dataset can reveal how different factors are associated with one another, hence facilitating the identification of unknown hidden patterns. This investigation will also present sufficient evidence to justify certain popular fallacies about terrorism [7]. One of the common misunderstandings is that a larger military can repress and control terrorism. Using the instrumental variable approach, however, studies demonstrate that counter-terrorism methods such as increasing military budget are insufficient to suppress terrorism, as it also depends on economic and political factors [8].

**Factors affecting the terrorism**

Identifying terrorism's dependant elements is one of the objectives of this study. There are characteristics, such as religion and nationalism, that are not defined in the dataset, yet have a significant impact on modern terrorism [11]. Religion has been a highly contentious topic among experts regarding whether religion influences terrorism and to what extent it does. Religious idealisation or belief can influence and change terrorism, according to conclusive evidence [12]. Religious idealisation has been one of the primary motivating reasons that have led to fanaticism and, subsequently, terrorism [13].

**Dataset challenges**

This dataset presents a significant problem due to the fact that individual investigations lead to contradictory outcomes. Current deficiencies and limits in data gathering methodologies, arguments over definitions, and irregularities in coding and processing give rise to disagreements among researchers, so invalidating their conclusions . A heuristic casual model showing relationships between globalisation and terrorism must be supported by an acceptable level of theoretical and empirical study . Critical disagreement over the definitional arguments surrounding major terrorist acts has a negative impact on the growth of this area . This issue necessitates exercising the requirement for common grounds that can be acknowledged by the majority of specialists and relevant authorities in order to agree on what may be the standard norms and method to be deemed a valid piece of information on terrorism upon which proper study can be conducted .

**DATA PREPROCESSING**

After data collection, the first stage is data preprocessing. It is a series of operations done on  dataset to change unclear data that can impede analytical conclusions. Raw data is essentially a collection of interconnected information. Oftentimes, raw data is unstructured and contains a great deal of information that is unnecessary to the project's requirements. Methodologies for data preparation facilitate the transformation of these raw data into a more meaningful, focused, interpretable, and readable manner.

The dataset from the Global Terrorism Database is insufficient, inconsistent, and contains several errors, missing attribute values, outliers, improper tags, and duplicate entries. These disparities can be resolved through data preparation.

The steps employed in this project's data preprocessing methodology are as follows:

**Data Cleaning**: Filling in missing numbers, removing outliers, and handling irregularities in data constitute the data cleaning process. Numerous categories, such as 'motives' and 'responsible organisations,' are absent from the terrorist dataset due to a lack of information or because the field was irrelevant to the incident in question.

**Data integration**: In this phase, data discrepancies are settled. To avoid misunderstanding and redundancy, different representations of the same data are combined, such as multiple subcategories of weapon type (weaponsubtype1, weaponsubtype2, and weaponsubtype3).

**Data transformation**: Here data aggregation, generalisation, and normalisation are conducted. The dataset contains several subtypes of target and victim. All of these subtypes were aggregated to reflect a single value by adding all subtypes with similar characteristics. This method minimises the overall number of attributes in the dataset, hence minimising the data's variability.

The dataset features a significant degree of data sparsity, which raises its overall dimensionality. This strategy diminishes the efficacy of density-related procedures such as clustering and outlier detection. Multiple fields include a greater number of missing or null values than legitimate ones.

**TECHNOLOGIES USED**

**Python** is an advanced programming language that supports multiple platforms, including Windows, Linux, Mac, and Raspberry Pi, among others. Python can be used to build online applications, database systems, manage large amounts of data, and conduct complicated mathematical calculations. Python is object-oriented, functional, and procedural. These are some of the Python packages utilised by this project.

**Mathplotlib** is a 2D plotting tool that includes the necessary modules and functions. A developer can adjust font properties, styles, axis properties, etc.

**Pandas** is used for data manipulation and analysis. Pandas can transform data structures and dataset formats to data frames on which operations like as loading data, renaming attributes, mapping, crosstab, sub-data frames, etc. can be carried out.

**NumPy** offers structures for multidimensional array objects as well as tools for associated operations. Typically, NumPy is employed for high-performance scientific computations.

**Seaborn**: It is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.

**plotly.express**  :The plotly.express module (usually imported as px) contains functions that can create entire figures at once, and is referred to as Plotly Express or PX. Plotly Express is a built-in part of the plotly library, and is the recommended starting point for creating most common figures.

Every Plotly Express function uses graph objects internally and returns a plotly.graph\_objects.Figure instance.

**Google Collab**: Collaboratory, or “Collab” for short, is **a product from Google Research**. Collab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

**Global Terrorism index;**

Using the describe function in selected data the indexes are:

1. This dataset consists of the Terrorists activities from the year: 1970:2017.
2. Maximum People killed:1570
3. Maximum People Wounded:8191
4. Total Casualties:9574

Using the statistical approach to deep dive into data to bring some insights:

Per Incident there is successfully execution of the Deadliest Attacks by perpetrators is

89%.

Out of which on per incident approx. 2.26 people are dying and 2,88 are injured.

*The arithmetic mean is the simple average, or sum of a series of numbers divided by the count of that series of numbers.*

*Standard deviation describes how dispersed a set of data is.*

And it might deviate from incident to incident to as shown in Std deviation column. This instances that in some incident there is huge loss of people and in some no loss.

|  |  |  |
| --- | --- | --- |
| **Indicator** | **Mean** | **Std Deviation** |
|  |  |  |
| Success (%) | 89% | 31.3% |
| Incidents | 1 | - |
| Killed | 2.26 | 11.22 |
| Wounded | 2.88 | 34.03 |

**Analysis**

1) Number of Attacks in each Year

The objective of this analysis is to better understand the terrorist activities through finding patterns and trends from the data. From this analysis, all activities that occurred in a given year in each available geographical location in the world are represented. This study allows us to determine whether terrorist activities are increasing or decreasing per year. We can see that in the year 2014 we had the most amount attacks and least amount attack in the year 1971. We can see that the number of attacks occurring increasing as the year progresses from 1970-2017 , so to understand the cause of drop in the number of attacks we made analysis further analysis on region wise attacks. So that we could see which region is worst affected and which is best at tackling terrorism.

2) Region wise Attacks

In the region wise attacks we focused on year and the region features to find how the distribution of attacks in the region across the years.

This distribution will helps us to better understand that which are the region have to face most of the attacks through out the years.

We can see that recently middle east and north Africa are the worst affected regions followed by South asia.

3) Success rates of Attacks by Terrorists group in each region

To understand how well the regions have been in controlling the situation and preventing the attacks from being successfully executed.

We can see that maximum attacks succesfully executed and failed by terrorists in Region Middle East & North Africa and minimum attacks successfully excuted was in region Australasia and Oceania.

4) Top 10 Terrorist Groups who has Successfully executed their attack mission

This analysis will tell us which are the most dangerous terrorist organisations in world. Here we are using features like terrorist group,success, number of attacks as our features to find the ranking of the terror organisation with respect to successfully executed missions

Terrorist Group:Farabundo Marti National Liberation Front (FMLN) has the highest success rate of their Attack mission with 3351 total attempts and Success rate 98.98%

Taliban is the Top terrorist group with the attck mission with 7478. It has the success rate of 89.32%.

5)Most attacked methods used by the Terrorist groups

Here from this analysis we are trying to understand which attack methods are favoured by the terrorist groups for their attack missions.

We select Attack\_Type and Terrorist\_Group as features to do this analysis.

We see that most of the Deadliest terrorist groups have preferred Bombing and Explosion for their Attack mission about around 50%(actual-48.6%). This means they have the target on masses and explosive sources available for their Deadliest missions.

6) Targets of the Deadliest Terrorists Groups

In this analysis we have considered features like Target, country, and terrorist group.

This was done to understand the which are the targets more prone to the terrorism organisations attacks.

Terrorist Groups has targeted the Private citizens and Property with 43.511k attacks followed by Military with 27.984k attacks.

**Correlation:**

**Correlation** refers to a process for establishing the relationships between two variables.

In Analysis no.7 we are using the Pearson method with heat map (to check the intensity of the correlation) and pairplot (using the scatter pairplot to visualize the linear relationship).

Pearson Correlation coefficient used for linear dependency between the data sets. The value of the coefficient lies between -1 to +1. When the coefficient comes down to zero, then the data is considered as not related. While, if we get the value of +1, then the data are positively correlated, and -1 has a negative correlation.

7) Correlation between Killed, Wounded and Total Casualty

In this analysis we have attempted to find the linear relationship using correlation method between killed, wounded and total casualty.

We have imported all the numerical features in our analysis data set to find the correlation. The correlation between wounded and Casualty is 0.99 means 99% of the variance in Casualty, is accounted for by the wounded. The correlation between Killed and Casualty is 0.80 means 80% of the variance in Casualty, is accounted for by the Killed.

The correlation between wounded and Killed is 0.69 means 69% of the variance in Killed, is accounted for by the wounded.

8) Total Number of Attack in Each Country

This analysis was carried out to see how each country in our dataset is affected by these attacks.

We have calculated the number of times the country has been there in the different terrorist attacks and through that we could find the total attack on each country.

We have plotted the top 10 countries with the highest attacks attempts.

**Conclusion**

That's it! We reached the end of our exercise.

Starting with loading the data so far we have checked the data shape, data head, data info. Then since the data contained lot of column labels so we filtered it to our need first then again imported the data to the collab. After that we looked for null values in the data set and started doing data cleaning till we successfully take care of null values. Then Data visualisation was done for the analysis to find different trends according to our need for better understanding.

**References**

[1] United Nations, “Chapter-1 Purposes and Principles,” [Online]. Available: https://www.un.org/en/sections/un-charter/chapter-i/index.html [Accessed: May 2019].

[2] A. Z. Borda, “Why we react differently to terror attacks depending on where they happen,” [Online]. Available: http://theconversation.com/why-we-reactdifferently-to-terror-attacks-depending-on-where-they-happen-57389 [Accessed: May 2019].

[3] START organization, “Global Terrorism Database,” [Online]. Available: https://www.start.umd.edu/gtd/about/ [Accessed: May 2019].

[4] ALMABETTER [Online],https://learn.almabetter.com/courses/take/team-capstone-projects/texts/19443222-world-bank-global-education-analysis