Capstone Project

EDA on Global Terrorism Analysis

(Kunika Gupta)
Data science students
Cohort- Hudson, Alma Better

Introduction

The **Global Terrorism Database (GTD)** is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The **GTD** includes systematic data on domestic as well as international terrorist incidents that have occurred during this period and now includes more than 180,000 attacks. 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. In this project, we are going to analyse the Global Terrorism Data to find out valuable insights.

Objective

The main objective of the analysis is to obtain the meaning full information and facts from the given huge datasets as shown above, by cleaning the datasets, doing proper analysis and visualization and plotting the useful information into different graphs and charts so that the trend and relationship between the various indicators on which the analysis is done can be understood easily.

1. Problem Statement

Data provided by the Global Terrorism Analysis (GTA) Data is in unformatted manner, corrupted data, and duplicate data and also sometimes it is irrelevant because it's piled-up data coming from various countries. For analyzing the data the data needs to be in the correct format and well-organized form.

The given data is as follows:

- **EDA GLOBAL TERRORISM ANALYSIS**

2. Steps involved

- **Loading and discovering data**
  Now, we need to load our data from the external source, which in this case is uploaded to the drive. The data is in the format of the CSV (Comma Separated Values) file.
- **Data Cleaning**
  Data cleaning is an important step in the data analytics process in which you either remove or update information that is incomplete or improperly formatted.

- **Null values Treatment by different methods**

  We have our dataset in hand which is raw and unfiltered. This step involves cleaning our data first by eliminating the columns which are not needed for our analysis. We have around 181691 rows × 135 columns in our dataset after removing the unnecessary columns we currently have around 16 columns with meaningful data that we could use in our analysis.

- **Exploratory Data Analysis**

  Exploratory Data Analysis is the approach of analyzing data, gathering and summarizing the important characteristics of the information, and using simple visualization that makes it easier to understand.

- **Importing necessary modules and libraries**

  We are importing the following libraries for their respective applications:

  Pandas:- Pandas is used to analyze data. It has functions for analyzing, cleaning, exploring, and manipulating data.

  Matplotlib:- Matplotlib is a graph plotting library in python that serves as a visualization utility. Most of the Matplotlib utilities lie under the pyplot submodule.

  Numpy:- NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices.

  Scipy:- This provides more utility functions for optimization, stats, and signal processing. Like NumPy, SciPy is open source so we can use it freely.

  Plotly:- The plotly is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.

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

  Datetime:- The Datetime module supplies classes for manipulating dates and times. While date and time arithmetic is supported, the focus of the implementation is on efficient attribute extraction for output formatting and manipulation.

- **Plotting various graphs for different parameters.**
- **Finding the key facts and relationships between various parameters.**
- **Observations according to the outputs of the graph visualizations.**

---

3. Data information

Global Terrorism Analysis (GTA):

Statistical information contained in the Global Terrorism Database is based on reports from a variety of open media sources. Information is not added to the GTD unless and until we have determined the sources are credible.

Characteristics of the GTD:

- Contains information on over 180,000 terrorist attacks.
- Currently the most comprehensive unclassified database on terrorist attacks in the world.
- Includes information on more than 88,000 bombings, 19,000 assassinations, and 11,000 kidnappings since 1970.
- Includes information on at least 45 variables for each case, with more recent incidents including information on more than 120 variables.
- More than 4,000,000 news articles and 25,000 news sources were reviewed to collect incident data from 1998 to 2017 alone.

Data visualization Analysis:

Data Visualization is the process of analyzing data in the form of graphs or maps, making it a lot easier to understand the trends or patterns in the data.

Correlation Heat map:

Analysis of the relation between the various columns of the cleaned data through the Correlation Heat map Matrix.

A correlation heat map is a graphical representation of a correlation matrix representing the correlation between different variables.

Observations:

From the correlation matrix above, we come to know that there is a strong correlation between the number of casualties and wounded people.

Word Cloud:

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud.

Observations:

As our word cloud shows that the frequency of the country Iraq is the most.

Geographical Heat map:

Geographic Heat Map is an interactive visualization that displays your data points on a real map and signifies areas of low and high density.

Observations:

The geographical heat map presentation of the total number of attacks that took place over 52 years of time.

Multi line Chart:

A multi-line chart is a basic line chart with one or more additional lines that represent comparison trends.

Observations:

We analyzed that attacks were taking place at nearly constant rate across the world for 4 decades, but in 2010 clearly a huge spike came in the Middle East and South Asia region. Although, it remained the same in other regions of the world.

Line Chart:

A line chart is a graphical representation of an asset's historical action that connects a series of data points with a continuous line.

Observations:

The line chart showed that the attacks were slowly rising from 1970 to 1990 and then a dip took place till 1998 and then a sudden spike in attacks were to be seen after 2010 across the world.

Pie Chart:

A Pie Chart is a type of chart in which a circle is divided into sectors that each represents a proportion of the whole.

Observations:

Iraq was the most attacked country across the world with 24636 attacks, followed by Pakistan, Afghanistan, and India.

Horizontal Bar Chart:

A horizontal bar chart is a graph in the form of rectangular bars. The length of these bars is proportional to the values they represent.

Observations:

Baghdad state which is the capital of Iraq is the most attacked in the world with 7645 attacks.

Donut Chart:

A donut chart is essentially a Pie Chart with an area of the center cut out.

Observations:

Firearms and Explosives are the most preferred attack types used by the deadliest gang Taliban. So, there should be stricter rules to prevent the movement of firearms and explosives into and from the countries.

Bar Chart:

A bar chart is a chart that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. A bar chart is sometimes called a column chart.

Observations:

The year 2014 had the most number of casualties, with the most number of attacks.

Stacked Bar Graph:

A stacked bar graph is a chart that uses bars to show comparisons between categories of data, but with the ability to break down and compare parts of a whole.

Observations:

By plotting region-wise attack data, we got to know that the success rate is very high in all regions. There is very less chance that an attempt to attack was missed out by terrorist groups. Also, we can see that the highest number of attacks happened in the Middle East and North Africa followed by South Asia, whereas Central Asia and East Asia are the least affected. Although in the Middle East and North Africa failed attacks are high, the percentage of that remains to be the same compared to other regions.

---

5. Summary:

- The year 2014 had the most number of terror attacks in the last decade. Approximately 17000 attacks in one year. This means that around 47 attacks were happening every single day during that year in multiple locations around the world.
-
- The Middle East & North Africa were the top affected region. 28.04 % of all events and a staggering 36.47 % of total casualties have been recorded from these regions.

- Iraq has been the country with the highest number of attacks and 21.87% of all casualties have been from Iraq.

- Baghdad has been the most attacked city in the world, 7 of the 26 cities are in Iraq and 4 of the 26 cities are from Pakistan.

- Bombing/Explosion has been consistently the most popular method of attack over the last 5 decades with 47.7% of all attacks.

- Taliban has gained much prominence since 2012 and is now responsible for the most number of terror attacks, and the most common attack type used by them are firearms and explosives. This has been the same for most terrorist groups.

- Around 82000+ attacks were done by unknown groups of terrorists, which is a major security concern.

- Private Citizens and Property are the most attacked targets, followed by the Military, Police, Government, Transportation etc.

8. Conclusion:

Starting with importing the data so far, we have done Exploratory Data Analysis, did null values treatment, and performed various types of data visualization techniques. Finally, from the analysis, we can conclude that **under-developed** countries are mostly getting targeted by terrorist organizations. **2014** was the most vulnerable year for

violence. Most attacks were happening in the **Middle East and North Africa Regions**, **Baghdad** which is the capital of **Iraq** was the most attacked state across the years. Most of the violent acts were executed by the **Taliban**. The most preferable weapon used by them were **bombs and explosives**.

References-

- **Global Terrorism Analysis(GTA):**

  https://www.start.umd.edu/gtd/

- **Python Pandas Documentation:**

  https://pandas.pydata.org/pandas-docs/stable

- **Python Numpy Documentation:**

  https://numpy.org/doc/

- **Python MatPlotLib Documentation:**

  https://matplotlib.org/stable/index.html