EDA on Global Terrorism Dataset

Soma Pavan Kumar
Data science trainee,
AlmaBetter, Bangalore

Abstract:

The Global Terrorism Database documents more than 180,000 international and domestic terror attacks that occurred worldwide from 1970 through 2017. With various dimensions of each attack. The GTD defines terrorist attacks as: The threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation.

Our Experiments will help us understand, what are the countries that were attacked the most and which cities were attacked the most. Also, which was the year that saw the most number of attacks. The types of attacks are mostly used by the terrorists and what type of targets are prefereed by the terrorists.

Keywords: Data Wrangling, Pandas, Matplotlib, Seaborn

1.Problem Statement

On The Given Global Terrorism Dataset, Explore and analyze the data to discover key findings pertaining to terrorist activities. And also, to find any relationships between dependent and independent variables.

2. Introduction

The Global Terrorism Database is an open-Source Database that includes information in terrorist attacks around the world from 1970 through 2017. This Database include systematic data on domestic as well as international terrorist attacks that have occurred during this time period and includes over 180,000 entries.

This 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.

The Global Terrorism Database (GTD) provides statistical information on terrorist events throughout the world starting in 1970 and is updated on an annual basis. Data is collected from news and media reports that are deemed credible by GTD investigators and researchers. The database contains information on more than 180,000 terrorist attacks.

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

3. Data Wrangling/Cleaning

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

Our Dataset includes more than 130 columns out of which more than 50 columns are empty and containing Data which are of no importance to our analysis.

So, we removed these columns one by one from the DataFrame using the drop function

We have used the pandas library which is an inbuilt library for all the data manipulations and data cleaning techniques.

Pandas:

It has functions for analysing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Pandas allows us to analyse big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science. Pandas are also able to delete rows that are not relevant, or contain wrong values, like empty or NULL values. This is called Data Cleaning.

Pandas library also provides us with functions such as describe() which provides us with statistics of the data including mean, mode, median and also Interquartile ranges.

Pandas library has some functions like shape, which gives us the number of rows and columns in the whole data.

Pandas also has some features like drop() which drops the whole column or a row as specified by the user.

4. Data Visualization

Data Visualization is an interdisciplinary field that deals with graphic representation of data. It is a particular efficient way of communication when the data is numerous as for example a time series

All of the Data Visualizations In this project are done using Matplotlib.pyplot, which is an inbuilt library in python that deals with plotting graphs including, bar graphs, histogram, all the way till pie charts and scatter plots.

Matplotlib :

Matplotlib is a low-level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter.

Matplotlib is open source and we can use it for free. Matplotlib is mostly written in python, a few segments are written in C, Objective-C and JavaScript for Platform compatibility.

Plotting in Matplotlib:

The plot () function is used to draw points (markers) in a diagram. By default, the plot () function draws a line from point to point.

The function takes parameters for specifying points in the diagram. Parameter 1 is an array containing the points on the **x-axis**. parameter 2 is an array containing the points on the **y-axis.**

In this project we have 5 main visualizations

1. The First one is to find the trend of number of terrorist attacks in a year and found out that year is 2014 is the year with max number of attacks.

2. The Second is to find the maximum number of attacks for a country and find out that Iraq stands first followed by Pakistan and Afghanistan.

3. The Third is to find the frequency of attacks according to their months and find out that there is no correlation between the number of months and attacks happening.

4. The fourth is to find out the maximum number of types of attacks used by the terrorists in their attacks to find out that Bombing/explosion is the most frequently used type of attack.

5. And, the last one was to find the most vulnerable targets of terror attacks and found out the obvious that people/ unsecure properties are the most vulnerable targets to terror attacks.

We also did data manipulation and found the top 5 cities that were attacked the most in all the years and the order is as follows:

1. Zinarag
2. Yokohama
3. Yacan
4. Wakunai
5. Vinchos.

Groupby Function:

A groupby operation involves some combination of splitting the object, applying a function, and combining the results. This can be used to group large amounts of data and compute operations on these groups.

Any groupby operation involves one of the following operations on the original object. They are −

- Splitting the Object
- Applying a function
- Combining the results

5. Conclusion:

The Dataset I choose to Explore is the data on Global Terrorism. The entries in the data range from 1970 till late 2017.

The Parameter that does not affect the number of terror attacks is the month, I.e Regardless which month of the year it is the number of terror attacks remain almost same and unaffected.

We also saw the preferred mode of attack for any terrorist is Bombing/Explosion and second on that list is Armed Assault(rifles)

And Terrorists attack private people more than any other target as they are unprotected and are also high profile targets(mostly individuals).
So, to Conclude, I Would say that this is a great Data to Explore and analyze on as a beginner.

One downside of this data is that most of the reasons why these attacks happen are political and cannot be determined by most Machine Learning Algorithms. But nonetheless it was fun to explore and analyze.

References:

- W3 schools
- Geeks for geeks