Data Visualization Assignment

Data Warriors

Justine Aeyels (r035274), Sanya Anand (r0823086),
Ethan Jones (r0827286) and Khachatur Papikyan
(r0825613)

supervised by
Prof. Jan Aerts

May 27, 2021

Introduction

The aim of the current project was to get acquainted with using data visualization as a tool to gain insights into large datasets. This project focuses on the visualization of the Global Terrorism Database.

Terrorism remains a global issue up until this day. Although it might seem that the threat for terrorist actions has decreased over recent years in Western countries, incidents as with fugitive military extremist Jürgen Connings show this threat never truly vanishes. As such, it was chosen to investigate the Global Terrorism Database, as terrorism is a contemporary and relevant topic of research. The Global Terrorism Database (GTD) is an open-source database downloaded from Kaggle (original source), containing information about over 180,000 terrorist attacks. Data visualization will make it possible to gain insight into this enormous dataset and to get a more clear overview of all past terrorist attacks. Three main research questions regarding terrorism were formulated, concerning the evolution of terrorism over time, its geographical dispersion and the difference between different terrorist organizations. Based on these questions the visual design process was started. Initially visualizations started out as basic sketches, which later evolved into intricate visualizations made using D3 and Tableau.

This paper first provides a brief description of the data set and variables of interest, then goes on to describe the design process that led to the final visualizations. To conclude, some insights which were acquired from the visualizations will be discussed.

Data Description

The GTD includes information on 181,691 terrorist attacks around the world from 1970 through 2017, except for the year 1993[1]. For an event to be included in the GTD as a terrorist act, the attack must satisfy the following definition: there must be a threatened or actual attack with the use of illegal force and violence to attain an either political, economic, religious, or social goal, and this by using fear, coercion, and intimidation. In practice this definition implies that attacks ought to be intentional, must involve some level of violence or threat of violence, and cannot be carried out by state actors, as acts of state terrorism are not included in the GTD.

The GTD consists of 135 features, of which 25 Numerical, 30 Textual, and

[1]Because of the challenges of retrospective data collection, only 15% of the incidents in 1993 could be identified having a source. Therefore, all attacks during 1993 were excluded from the GTD to avoid the misinterpretations of low frequency of events that year.

80 Categorical. For each attack, these features include information such as its date, location, weapons used, nature of target and number of casualties. 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.

There are a total of 135 features included in the dataset, which can be further categorized in following categories: ID, incident date, incident location, incident information, attack information, target/victim information, perpetrator information, perpetrator statistics, claims of responsibility, weapon information, casualty information, consequences, kidnapping/hostage taking information, additional information, and source information. Further investigation of these 135 features found that some of these are incomplete and/or redundant. Hence, only 31 columns that seem to be of greater interest in the context of further analysis are summarized by their descriptions and some infographics in Table 1 below. As the analysis proceeds and new insights and questions arise, some remaining variables might still be included in the final visualizations.

To get a feel of the dataset, some graphs are provided displaying some distributions of the most important variables in the dataset in various ways (Figure 1). More detailed information about variables can be found here.

Figure 1: Exploratory data visualizations

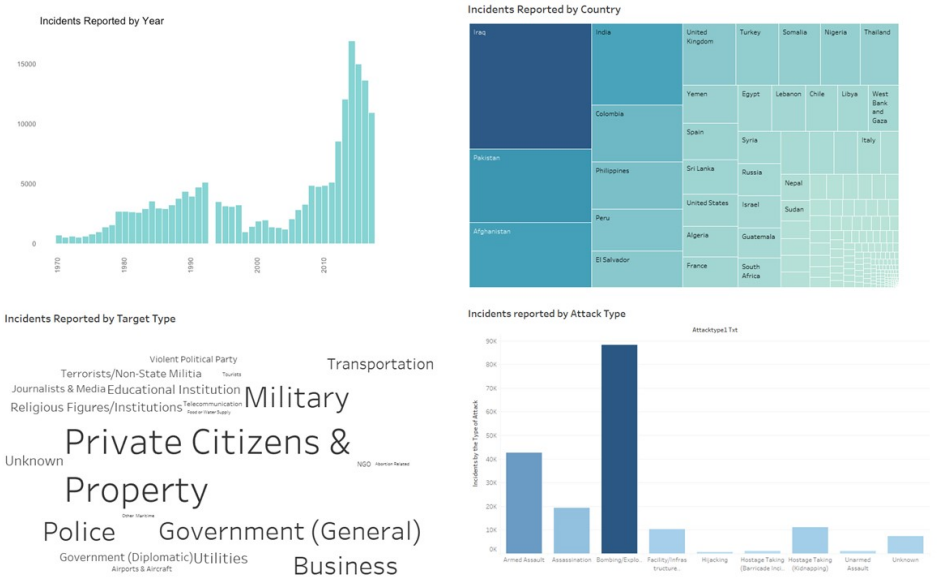


Table 1: Columns of key interest

Name	Description	Type
eventid	A 12-digit Event ID system. First 8 numbers – date recorded "yyyymmdd". Last 4 numbers – sequential case number for the given day (0001, 0002 etc).	Numeric
iyear, imonth, iday	Numeric year, month and day when the incident occured.	Numeric
extended	The duration of an incident extended more than 24 hours: 1 = "Yes" , 0 = "No"	Categorical
country, country_txt	The country code and the country or location where the incident occurred.	Categorical
region, region_txt, provstate	The region code based on 12 regions, the region name in which the incident occurred and the name (at the time of event) of the 1st order subnational administrative region	Categorical
longitude, latitude	The longitude and the latitude of the city in which the event occurred.	Numeric

summary	Textual summary of the incident. Contains 36% missing values.	Text
success	Success of a terrorist strike: 1 = "Yes" , 0 = "No"	Categorical
suicide	1 = "Yes" The incident was a suicide attack. 0 = "No" There is no indication that the incident was a suicide attack.	Categorical
attacktype1_txt	The general method of attack and broad class of tactics used.	Categorical
targtype1_txt, targsubtype1_txt	General and more specific information on the type of target/victim	Categorical
target1, natlty1_txt	The specific person, building, installation that was targeted and/or victimized and the nationality of the target.	Categorical
gname	The name of the group that carried out the attack.	Numeric
motive	Textual description of the motive for the attack. 72% of the values are missing, still may be interesting.	Text
nperps, nkillter, nwoundte	The total number of terrorists participating in the incident, the number of killed and wounded among them.	Numeric
weaptype1_txt weapsubtype1_txt	General type of weapon used in the incident and more specific value for most of the weapon types.	Categorical
nkill, nwound	The number of total confirmed fatalities for the incident and the number of wounded.	Numeric
property, propextent	Property damage and the extent of it.	Categorical

Research questions

By Region

In the BBC article, "The Changing Faces of Terrorism", Robert Adams wrote terrorism tends to become "endemic" to a region, where "started in a nationalist cause, it is then employed in resistance to the resulting state. Started

to cleanse society of corruption and external control, it continues in support of the drug trade and prostitution. If violence becomes a habit, its net effect can be to prevent economic development, to provide a justification for official violence, and to perpetuate existing patterns of dominance and submission."

By taking an "endemic" perspective on terrorism, this research question investigates the regionality of terror actions. Does the violence tend to be evenly distributed among the regions or is it only dominated in a few regions? Do the methods of attacks and targets differ between regions? How are different regions affected in terms of property value as to human life. Extending this regional perspective to account for geography, we may investigate if there are paths terrorists take and what drives them towards that path — resources of the region, ease of logistics, legislation, socio-cultural level of ethnic population, or religion?

Between Organizations

This research question investigates patterns of influence between terror organizations. To this end, visualizations assess groups that share naming conventions, as this best reflects how an organization identifies itself and associates with others. By categorizing terror organizations' patterns in naming, is there further insight into a pattern of influence among groups?

Over Time

Adam Roberts, Ph.D., provides a narrative of the broadening targets for terrorist events in recent years, contrasting targeted political assassinations in the late 19th and early 20th-century, to the late 20th century where he describes the prevalence of organizations to "inflict carnage and destruction on their adversaries." A similar sentiment is echoed in the last chapter of Revill's "Improvised Explosive Devices: The Paradigmatic Weapon of New Wars," where he claims that IEDs have existed throughout the modern era, but have "come of age" in the 21st-century due to advancements and availability of necessary components, and now demonstrate "increasing volume and deadly efficiency."

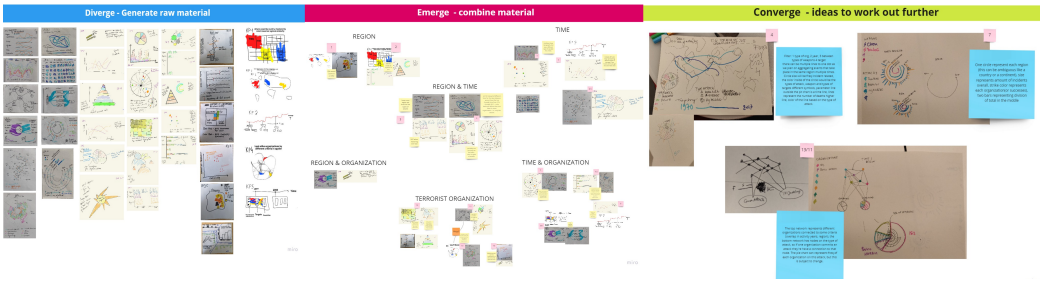
This leads to an inquiry on how terrorism has changed over time: Has there been an evolution in the methods of attacks or has there been any prominent type of attack that emerged over time while considering the advancement in technology? Has any terrorist organization become more active over the years? Can we find any seasonal or longitudinal patterns in terms of attacks, weapons used, and most importantly targets?

Exploration of the Design Space

Exploration of the design space started off with everyone individually drawing a multitude of sketches in which various aspects of the dataset were visualized. The book Gamestorming (2010, Macanufo & Brown) refers to this process as the diverge phase. Once a sufficient amount of sketches were gathered, each of these were discussed in group, after which they were clustered together depending on the research questions they provided insight into, i.e. the emerge phase. Six main clustered were identified: visualizations providing intel about how terrorism evolved over time, how terrorism manifested within a region and how terrorism differs across terrorist organizations, but some visualizations already offered insights into more than one research question; i.e. about both the time and spatial dimension of terrorism, about the evolution over of terrorist organisations, and about the geographical distribution of these organizations. The last step in this exploration process, also called the converge phase, was to gather the most insightful visualizations from each cluster and try to merge them into new visualizations which gave an even more in depth view into the data.

Below, Figure 2 shows the full miro board, outlining these phases under their respective titles "diverge", "emerge" and "converge". Each of these steps will now be discussed in more detail for each research question. It should be mentioned however that these final sketches are provisionary, as during the implementation phase new ideas emerged for adapted version of these sketches or entire new visualizations.

Figure 2: Exploration of the Design Space in Miro

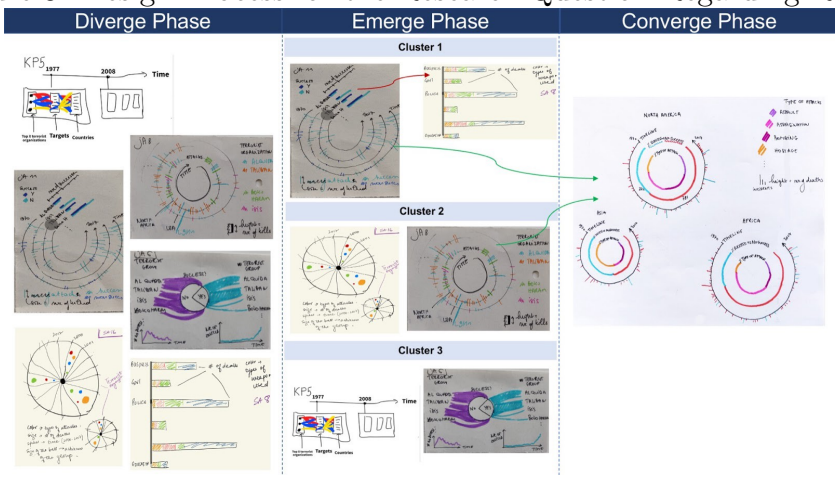


By Region

An overview of the main steps of the design process for the research question regarding the geographical distribution of terrorism is displayed in Figure 3.

In the final visualization (Figure 4) different circles are displayed for each region, each consisting of different rings portraying a different aspect of the attacks in that region. Circle size here represents the overall amount of incidents by region, with larger circles representing regions a higher amount of incidents. The outside ring of the circle represents a time-scale (from the beginning to the end of the data set) with strikes representing an incident in that region. The height of the strikes represents the number of casualties for the specific attack and the color represents whether the attack was classified as a success or a failure. The ring below the time-outline represents the cumulative amount of the successful versus unsuccessful incidents in the region, using the same color code as for the strikes on the time-outline. Similarly, the last ring shows the cumulative distribution of the different types of attacks over all incidents in the region.

Figure 3: Design Process for the Research Question Regarding Region

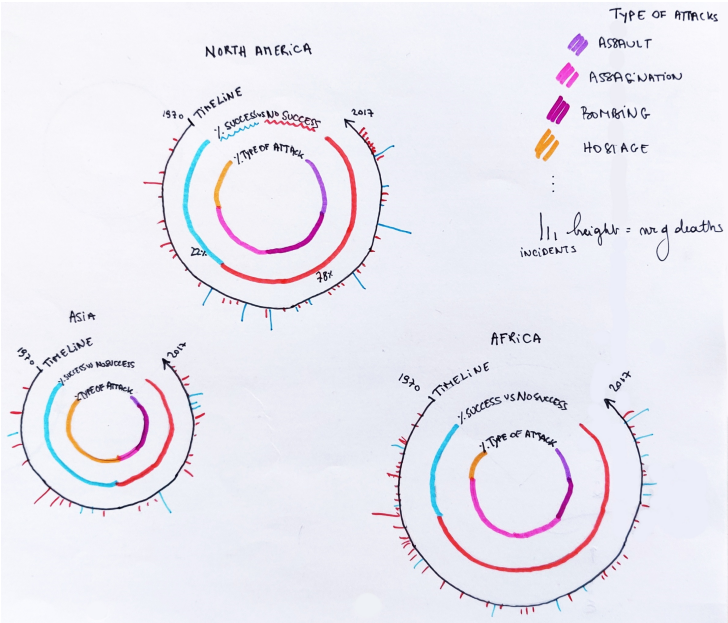


Between Organizations

The design process leading to the final visualization (Figure 6) regarding the between-group comparison of terrorist organizations is displayed in Figure 5.

The image is divided into two networks: the top network connects similar terror organizations, and the bottom network connects organizations to a similar attack type. On the top network, nodes represent different organizations combined in a network based on a relational criterion. These criteria have yet to be finalized but can represent a variety of concepts (i.e. the regional proximity of organizations, overlap in years active, motivations of each organization, etc.) The bottom network disseminates information on

Figure 4: Final Visualization for the Research Question Regarding Region



the type of attack. Each organization would be present in a hybrid pie/bar chart showing the contribution of each to the total casualties of the type of attack.

Over Time

Again, Figure 7 shows the design process that led to the final sketch (Figure 8) regarding the by time research.

It was chosen to use a map representation to describe the evolution of terrorist attacks through time around the world. To ensure the map remains legible, it was chosen to display this evolution separately for different terrorist organizations, for which a filter will be available to select a specific terrorist organization. On this map, each circle represents a region, connected by a line to the area where the next attack occurred. The line color will darken from past to present showing the regional spread of incidents over time. There can be multiple lines going to the same circle, (for example the line can go back and forth between two regions multiple times) but the line won't emerge from a region to represent consecutive attacks in that same region, the line only exists when the next attack happens in a new region.

Circle size represents the frequency of incidents, with larger circles indicating a higher number of incidents in the region, and inside the circle, a pie chart shows the makeup of different types of attacks in the region. Sim-

Figure 5: Design Process for the Research Question Regarding Terrorist Organizations

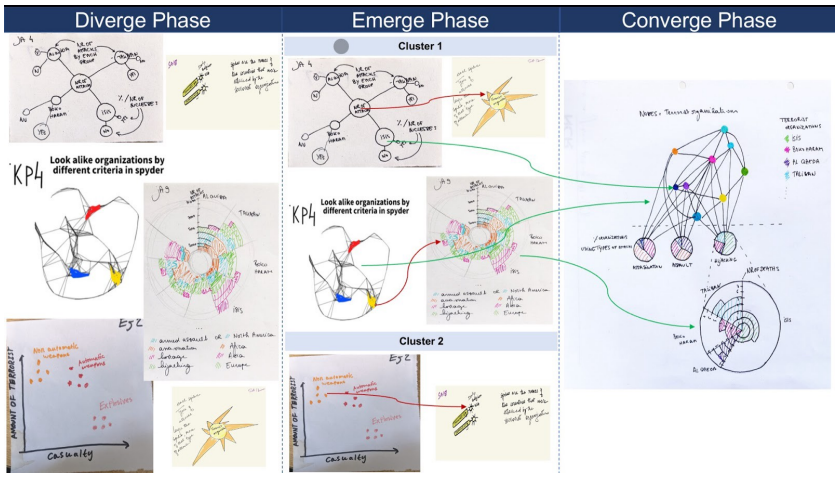
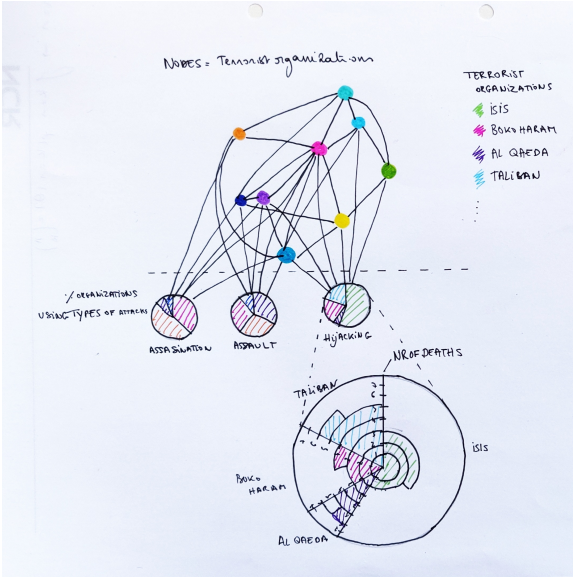


Figure 6: Final Visualization for the Research Question Regarding Terrorist Organizations



ilar to the previous drawing, the circle's perimeter is a timeline with strikes representing terrorist attacks, casualties (height), and type (color) for each attack.

Figure 7: Design Process for the Research Question Regarding Time Evolution

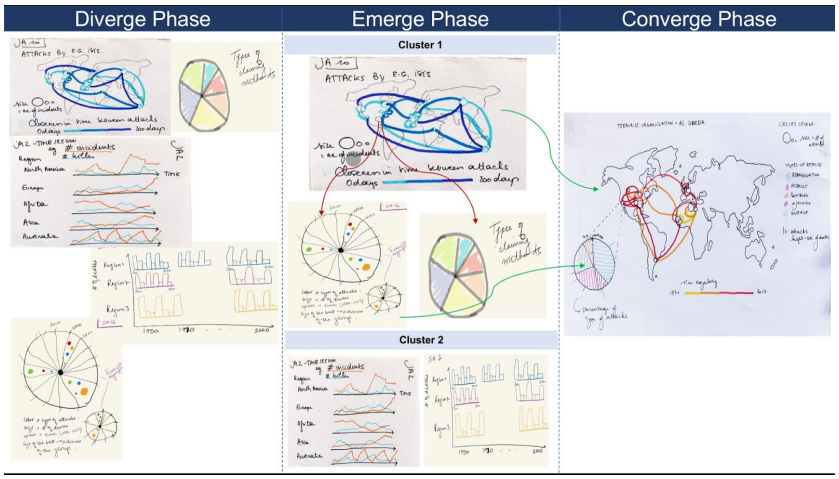
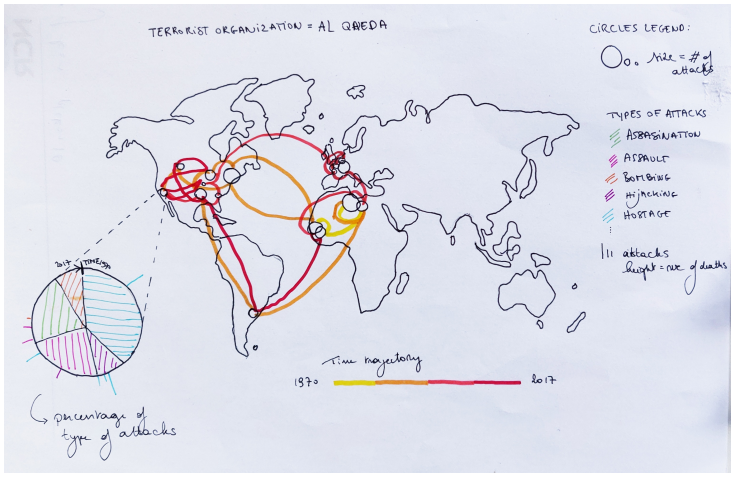


Figure 8: Final Visualization for the Research Question Regarding Time Evolution



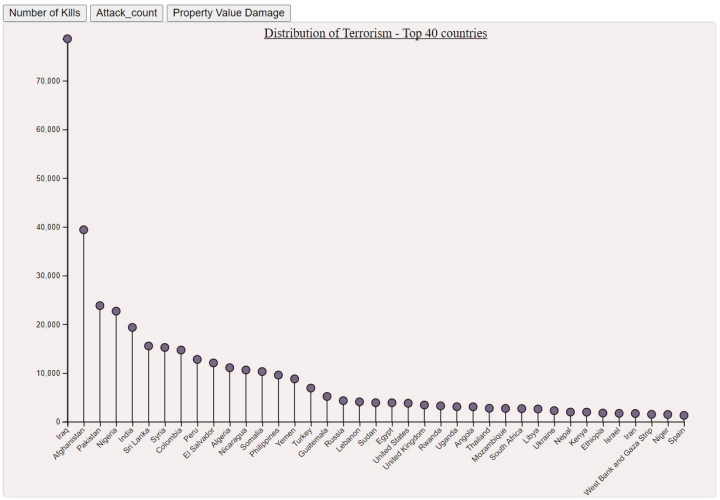
Final Designs

By Region

A total of four visualizations were eventually made with regard to the research question looking at terorism over time. Three of these will be briefly discussed below. Because of the constraint on the number of pages the last one is included in the Appendix and will not be discussed in detail further.

First off, Figure 9 displays a lollipop chart that looks into the differences between 40 countries in regard to the numbers of terrorist incidents, the number of kills due to these attacks, and the total property value damages caused by the attacks. The buttons above the charts allow to switch between the different charts, which makes it easy to compare the countries on these different criteria. An additional tooltips provides more detailed information about the visualization.

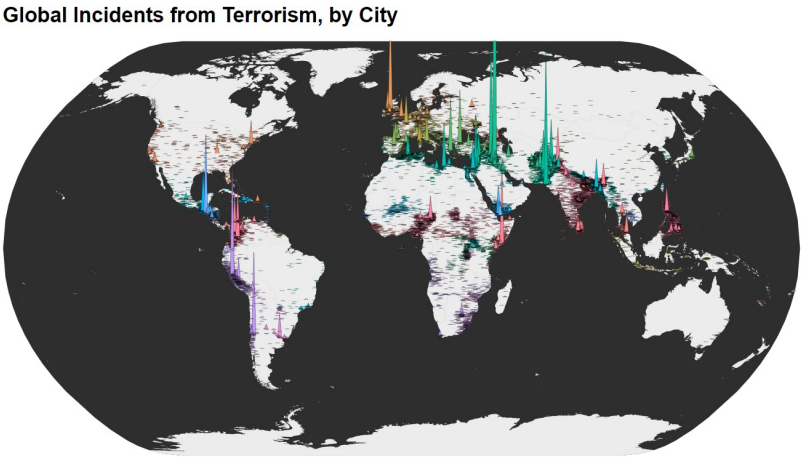
Figure 9: Lollipop Chart



Next, a mountain chart (Figure 10) visualizes all terrorist attacks that occured in different cities around the world. Peaks display the cumulative attack in a city, with higher peaks representing more targeted cities. Each peak is interactive, once scrolled over the border stroke increases and a tooltip displaying the name of the city and number of incidents appears. The chart's also pannable, allowing the user to zoom in on locations. Colors differentiate based on the country each peak belongs to.

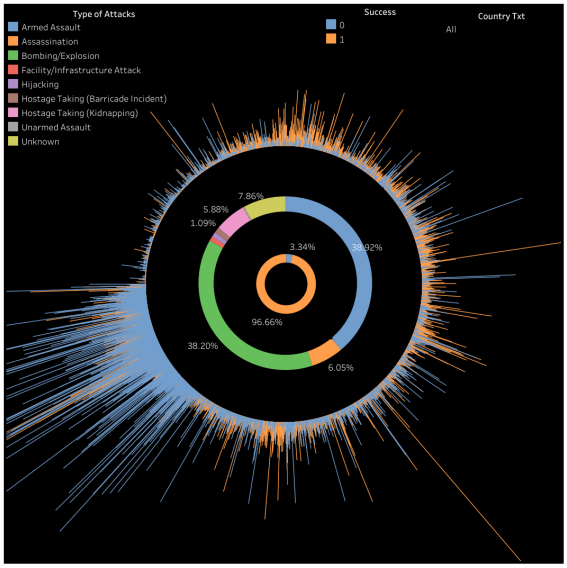
The third visualization is a radial donut chart (Figure 11), which compactly visualizes attack history, with the option to subdivide by country.

Figure 10: Mountain Chart



The inner and middle donut charts represent the percentage of success versus failure of attacks, and type of attacks, respectively. The outer layer is a bar chart chronologically measuring the number of deaths from each action, where bar height is relative to the number of deaths caused by said action in the country. According to the codebook, success of a terrorist strike is defined according to the tangible effects of the attack and is relative to the type of attack. This provides a qualitative perspective of the attack's execution, events can be considered unsuccessful despite a high death toll.

Figure 11: Radial Donut Chart



Between Organizations

By-group visualizations display associations between groups that share similar naming conventions. The methodology is addressed further in the implementation section, but at a high level, the analysis operates under the assumption groups would name themselves similarly to groups they identify with or are inspired by. By subdividing groups that incorporate the most popular naming conventions, the charts look to compare groups within and between different naming patterns, with Figure representing the latter. The Sankey chart (Figure 12) organizes this naming categorization through time and region. Each terror organization (falling among the seven naming conventions of the left) is charted once through the decade and region of their first attack. Interactivity provides a tooltip on the number of groups from one node to the next. Figure 13 represents a "within-group" dimension, supplementing the Sankey by displaying actions of individual groups within similar naming conventions. The origin point (0,0) represents the first act committed by a group within that naming convention, with dots representing all actions by organizations in that group. Unlike Sankey, the same organization can be responsible for multiple visualizations. The Y-axis resembles distance in kilometers and X-axis distance in a time of subsequent attacks from this origin point. The interactive color legend (as shown on the right) clarifies the action's region.

Over Time

The following Figure 14 is a still frame from an animated chart displaying yearly attacks based on the location they occur. The blue dots represent deaths from attacks, with size increasing between large and small amounts of casualties. Red dots show terrorist attacks of the previous years and remain at lower opacity for each year, fading after a total of three years from occurring. The map can be viewed both in a static and dynamic way. Specific years can be selected or an animation can be started which subsequently displays the attacks throughout the years, starting at 1970 until 2017, at any chosen speed.

Different filters are available for users to filter which attack type to display. Though dot size displays deaths from terror incidents more precise filtering is available through setting the "Nkill" (i.e. number of deaths due to the terrorist incidents) filter to the desired amount. The chart is also available for specified regions (Figure 15).

Figure 12: Sankey

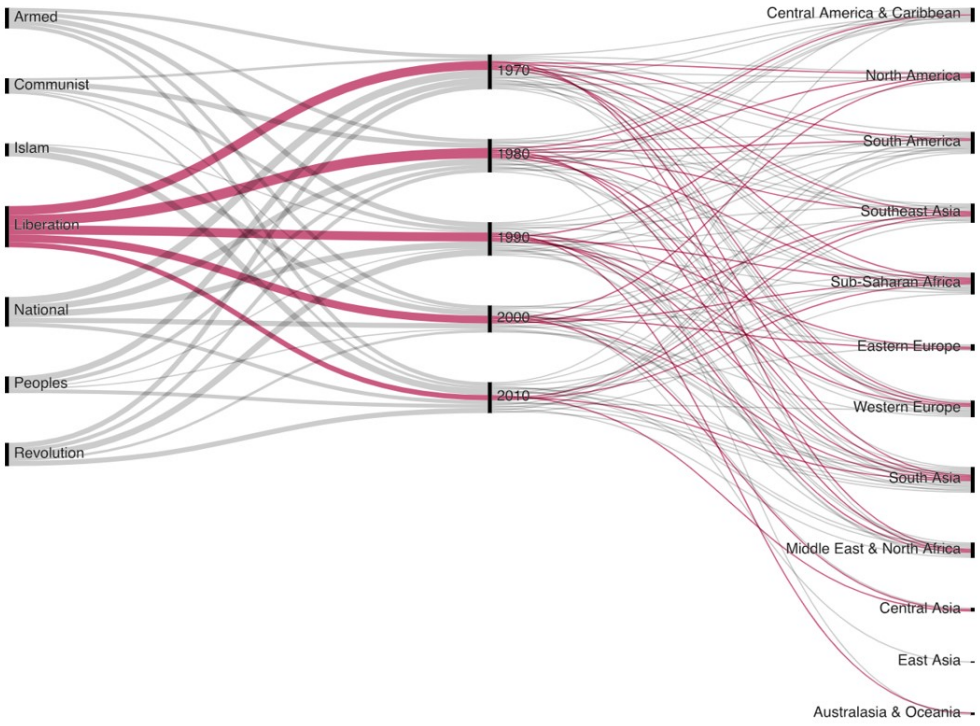


Figure 13: Scatterplot

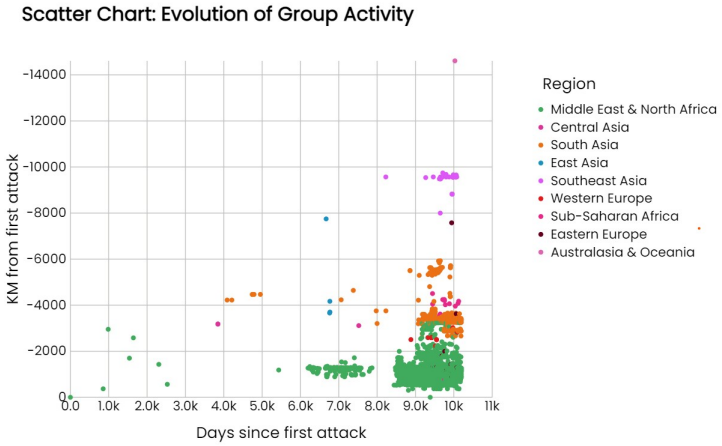


Figure 14: Evolution of Terrorism over Time Map

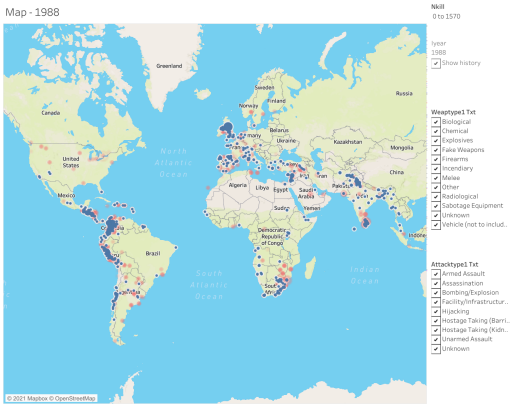
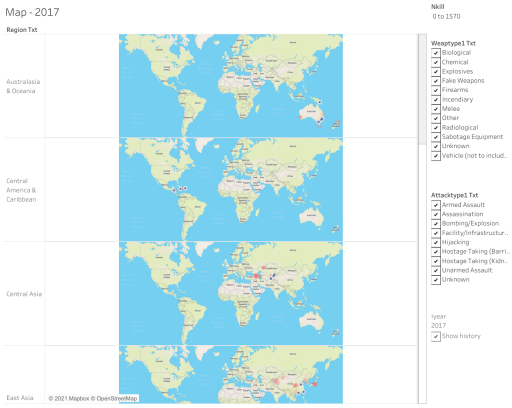


Figure 15: Evolution of Terrorism over Time Map for each Region



Implementation

Implementation took place in three phases: data preprocessing/wrangling, visualization, and hosting. The following briefly summarizes aspects of preprocessing and technical implementation for each visualization. The section is concluded with a brief description of the shared hosting method.

By Region

The original dataset was cleaned by removing columns with more than 85% missing data removed and filling all other NAs items with "0". For Figure 9 data is organized by cumulative country totals of people killed in attacks, cost of property damage (in United States Dollars), and the number of terror attacks. Only the top 40 countries were considered for this chart based on

the highest number of killings. Figure 10 aggregated the total number of incidents by city. Both charts were visualized using D3; each implementation consists of separate CSS, Javascript, and HTML files.

Figure 11 is created through Tableau. This allowed the data to be visualized without prior modification, apart from using a cleaned dataset. The radial tick chart in the perimeter of the visualization is created through a customized data mapping: radial angels and X, Y coordinates were calculated to display data in the intended format. In a dashboard, the radial bar chart was combined with two donut charts representing attack type and proportion of killings for success/non-success label. Finally, a drop-down was created to allow users to filter data by country.

For Figure X (World Chart), a subset of data was considered to include only the 25 countries with the highest terrorist-related killings and was visualised using D3

Between Organizations

The data was preprocessed using Python and R. The 16 most popular two-word combinations in organization names were used to subset and categorize the data. These two-word combinations are then further categorized by their starting words. Starting words represent the bucket of popular two-word combinations. They do not represent every terror group containing that word. Word combinations are shown in Table 2 below.

In instances where an organization might belong to multiple different naming categories, the group was assigned the most popular category. For example, a group with the name "People's Liberation Movement" falls under "People's Liberation" and "Liberation Organization", but would be filled under the latter because it's the more popular two-word combination. This is the basis of the text categorization driving the visualization in Figures 12 and Figure 13.

Figure 12 is inspired by Mike Bolstocke's parallel chart visualization, initially used for comparing two groups across different nodes. The use of an end-to-end color selection of a single category led to the adoption of the chart for this problem. The data were processed in R to match the correct format. Figureref 13 is an adaptation from a scatter plot by Curran Kelleher's which utilizes D3 and React to create an interactive legend.

Figures 12 and Figure 13 are combined on a shared webpage, with 7 web pages for each word category. This implementation is selected because it facilitates better organization of data and code.

Table 2: XXX

Name Type	Description
Liberation	“liberation army”, ”liberation movement’, “liberation organization”
National	“national socialist”, “national liberation”
Communist	“communist party”
Armed	“armed forces”
People’s	“people’s revolutionary” “people’s army”
Islam	“islamic state” ”islamic movement”
Revolution	“revolutionary army”, “revolutionary movement”, “revolutionary action”, “revolutionary party”, “revolutionary armed”

Over Time

Similar to the visual radial chart, Figure 14 and 15 are created in Tableau using the cleaned dataset. The implementation visualized dots, where size represented the number of kills at the latitude and longitude of the event, over a year. Additionally, dots in red and lower opacity are displayed for up to three years prior. The default map template was replaced with a template where countries are clearly labeled.

The visualization is hosted on Tableau public. Filtering methods were added to allow users to filter by attack type, weapon type and the number of kills occurring in a single event. An animation effect is created by visualizing charts sequentially by year. A modified version of the visualization (Figure 15) uses the same world map template, but subset the data to specific regions to allow clarity for terrorist activity in different areas.

Hosting

A Tableau public profile is created to host all Data Warriors visualizations in Tableau (Figure 11 and 14). The other visualizations are hosted on GitHub as separate websites. Figures 12 and 13 are displayed on the same website, with web pages for each visualization, and Figure 9, Figure 10 and Figure 17 have separate web pages.

Insights

By Region

Terrorism is often regionally focused but we can infer from the Figure 9 that terrorist attacks are heavily concentrated in few countries only. The Middle East, Africa, or Asia are the regions with the maximum number of deaths but not all countries in these regions were affected. Iraq is the country with the most assaults and hence, has the maximum number of fatalities. Interestingly, it is the country with the lowest property value damage. The intensity of attacks in the UK is severe in terms of property value damage and surprisingly with very few terror attacks in comparison to other countries.

As can be seen in Figure 10 Iraq is the country where most terrorist incidents emerged and primarily in the cities of Baghdad and Mosul. The tall peaks in Pakistan, UK, and European Union indicate these attacks tend to occur in targeted cities. In contrast, terror actions in other countries are more dispersed over the region. Some of the dispersions have the effect of making a dark trail on the map due to very few events. These dark trails

resemble areas where organizations have resources to rival regional control. Instead of "one-off" terror attacks, these instead represent campaigns with multiple conflicts. Such is the case with the Maoist Communist Party of India in east India. Table 2 presents a comparison of visualization by Al Jazeera of the organization's activity.

Figure 11 reveals the extent of attacks in the U.S. has been very low, but has spiked with the 9/11 attacks, which is considered the world's most fatal terrorist event of all times. It claimed over 3000 lives and the total deaths alone in 2001 is almost four times higher than the total killings from terrorism in the U.S. since 1970. According to the dataset, the incidents of hijacking make less than 1% of deaths but over 70% belong to the U.S. , and almost half of the deaths came from 9/11. Most of the major terrorist attacks such as Madrid train bombings -2004, 2005 London bombings, 26/11 Mumbai attacks - 2008, 2015 Paris attacks, and others have occurred within 15 years. Also, the majority of the assualts came from bombings/explosions armed attacks and over 65% of the attacks in Iraq, the most terrorist attacked country, were from bombings/explosions.

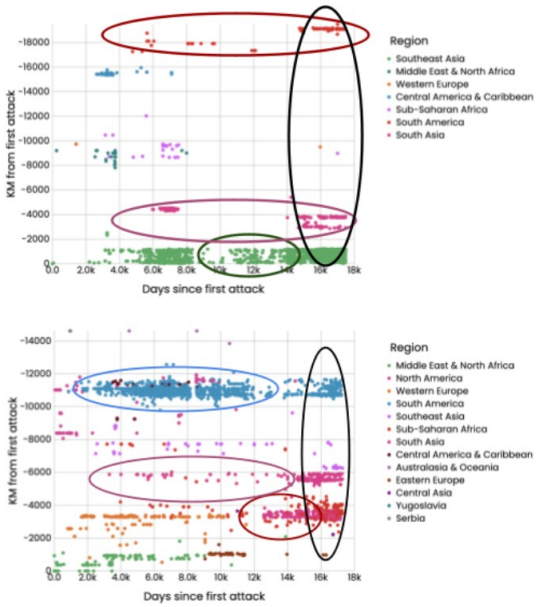
Between Organizations

A pattern of resurgence is observed across multiple group categories (see Figure 16. The pattern can be described as a region with a past of terrorist activity under one of these naming categories, experiences further increased activity, when actions of groups under the same naming category are occurring elsewhere. This pattern is demonstrated in annotated graphs for organizations under the "People's" (top) and "Liberation" (bottom) naming categories, where long horizontal ovals demonstrate a region's past activity, a small horizontal oval highlights an increase in activity, and the vertical black oval the resulting increase in activity. This could be the result of increased activity in one region causing another region to recall its history. This reflects, if not by new organizations being formed, organizations doing actions under new more popular names

Over Time

An interesting time pattern can be detected when examining the over time developments of terrorist attacks in Figure 14. Up until a certain moment in time visually there seems to be no terrorist attacks in the territories of the former Soviet republics, while in the other parts of the world the attacks were evenly distributed. From a bird's-eye view this could be interpreted as an insight in terms of changes in the over time trajectory of terrorist attacks

Figure 16: Insights regarding the between organization research question



stating that after the collapse of Soviet Union the post Soviet countries became interesting destinations for terrorist organizations. However, this kind of conclusions should be treated with caution, because there could be a data collection issue meaning that the events behind the iron curtain of the USSR might not be reported to the general public and as such, not included in the current dataset as well.

Contributions

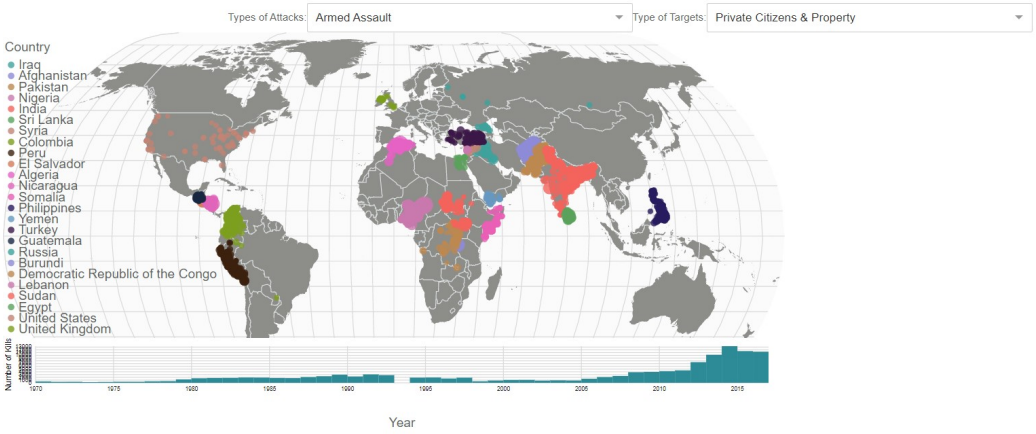
Up until implementation everyone contributed equally to the assignment. The actual visualizations were divided among us: the visualization for the research question regarding the between-group comparison was done by Ethan, the region research questions mainly by Sanya, and the over time evolution by mainly Khachatur and Justine.

Appendices

Appendix

Additional visualization (Figure 17 for the by-region research question

Figure 17: World map



Links

- GTD: Kaggle and Original Source
- Miro Board: link
- Video: link
- Figure 9: visualization and code
- Figure 10: visualization and code
- Figure 11: visualization and code
- Figure 12: visualization and code
- Figure 13: visualization and code
- Figure 14: visualization and code
- Figure 15: visualization and code
- Figure 17: visualization and code

List of Figures