Contents

1 Introduction 4

2 Record layout 5

3 Schema 6

3.1 Analysis 7

3.1.1 Tree editing 9

3.2 Final tree 11

3.2.1 Dimensions and fact schema 12

3.3 Fact schema 14

4 Data cleaning and preprocessing 15

4.1 Extraction, Cleaning and Transformation 15

4.2 Load 18

4.3 Data quality 20

5 Analysis sheets and Dashboards 22

5.1 USA dashboard 22

5.2 Terrorists dashboard 23

5.3 Middle East and Africa dashboard 25

5.4 Deaths and injured dashboard 26

6 Conclusions 28

List of Figures

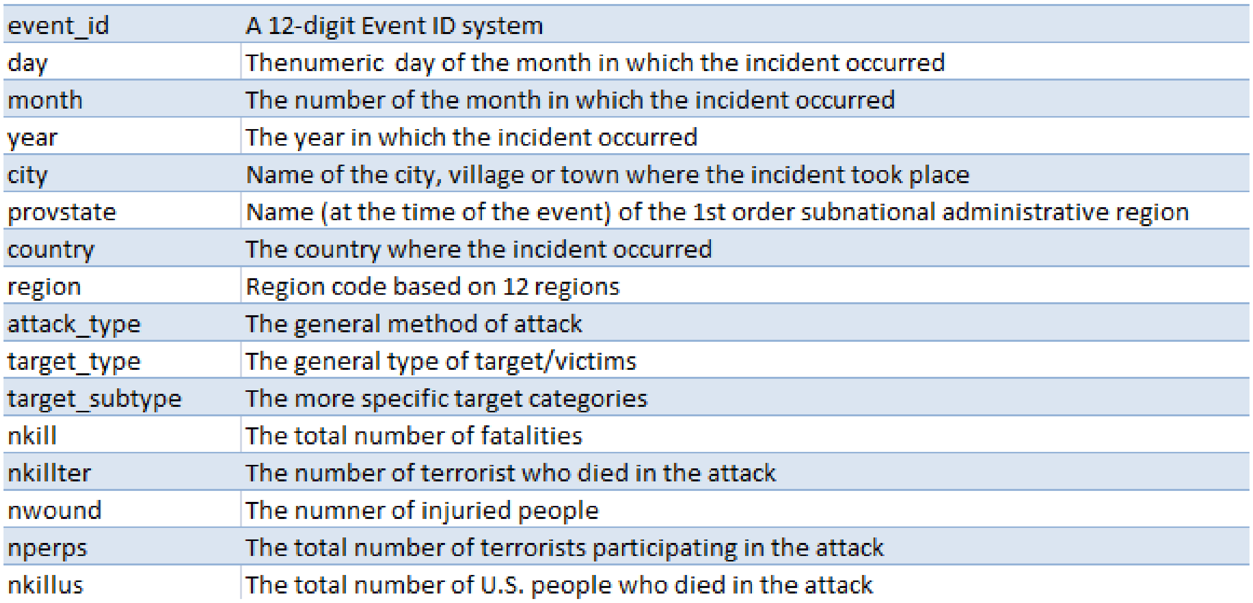
1 Record layout 5
2 Attribute tree 6
3 Tree cuts pt. 1 9
4 Tree cuts pt. 2 10
5 Final Tree 11
6 Final schema 13
7 Fact schema 14
8 Cleaning and transformation 17
9 Date load 18
10 Event 19
11 Data quality 21
12 Dashboard USA 22
13 Dashboard terrorists 24
14 Dashboard middle East and Africa 25
15 Dashboard deaths 26

1 Introduction

The analyzed data comes from the Global Terrorism Database (GTD), which 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 time 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. The data set was found on Kaggle (link to it) and it provides us with a lot of information about over 180.000 attacks. Every worldwide attack between 1970 and 2017 (except for 1993) is reported into the data set, telling us the country where the attack took place, the number of deaths, the attack type etc... for a total of 135 columns.

2 Record layout

The most important attributes, and which will be taken into account in the analysis of the data set, are summarized in this table:



event_id	A 12-digit Event ID system
day	Thenumeric day of the month in which the incident occurred
month	The number of the month in which the incident occurred
year	The year in which the incident occurred
city	Name of the city, village or town where the incident took place
provstate	Name (at the time of the event) of the 1st order subnational administrative region
country	The country where the incident occurred
region	Region code based on 12 regions
attack_type	The general method of attack
target_type	The general type of target/victims
target_subtype	The more specific target categories
nkill	The total number of fatalities
nkillter	The number of terrorist who died in the attack
nwound	The numner of injuried people
nperps	The total number of terrorists participating in the attack
nkillus	The total number of U.S. people who died in the attack

Figure 1: Record layout

3 Schema

As first thing, after making the relational model, I started making the attribute tree and then, after choosing the analysis that I wanted to carry out, I made the cuts on the tree, eliminating the attributes that weren't useful.

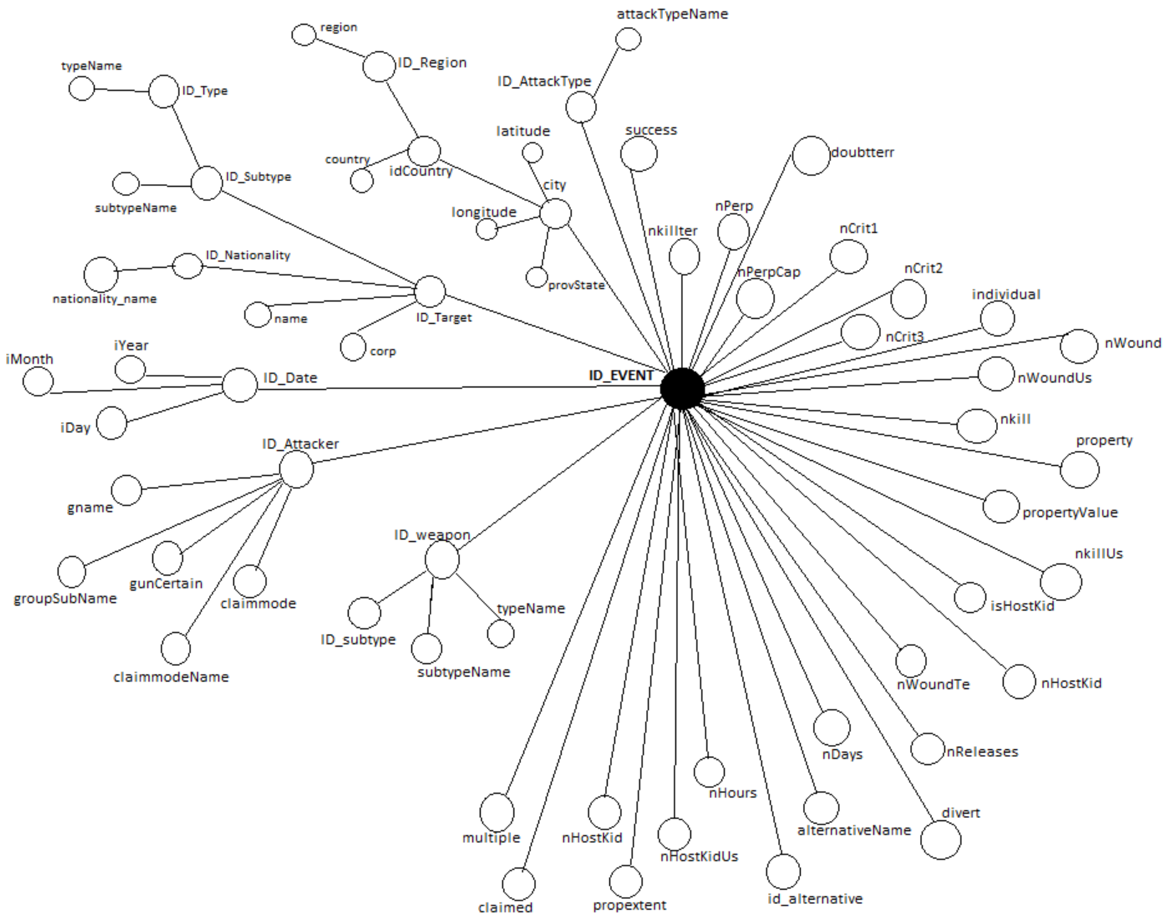


Figure 2: Attribute tree

3.1 Analysis

On the data set I've chosen to do the following analysis:

- A comparison between the number of deaths and the number of injuries per year.
- How many terrorist participated in every attack and how much of them died in it, showing the percentage of how many of them survived.
- A comparison of the target of the attack (clinics, royalty, legal services airport, etc) and the type of the attack (armed assault, bombing, hijacking, etc), counting how many occurrences by association there have been in the whole world.
- The number of deaths for every country since 1970.
- The trend of the injured in Africa per year.
- A general annual trend of the total number of attacks, and an estimate is also made based on this trend, which predicts (grossly) the number of terrorist attacks up to 2023.
- A comparison of the total number of attacks per year between middle-east and north Africa, south Asia and sub-Saharan Africa.

- The number of attacks by type of them and making a comparison of them between Africa and USA.
- A comparison of the US deaths by attack type
- The number of US people that died in attacks all over the world
- The percentage for year of US people that died against all deaths

3.1.1 Tree editing

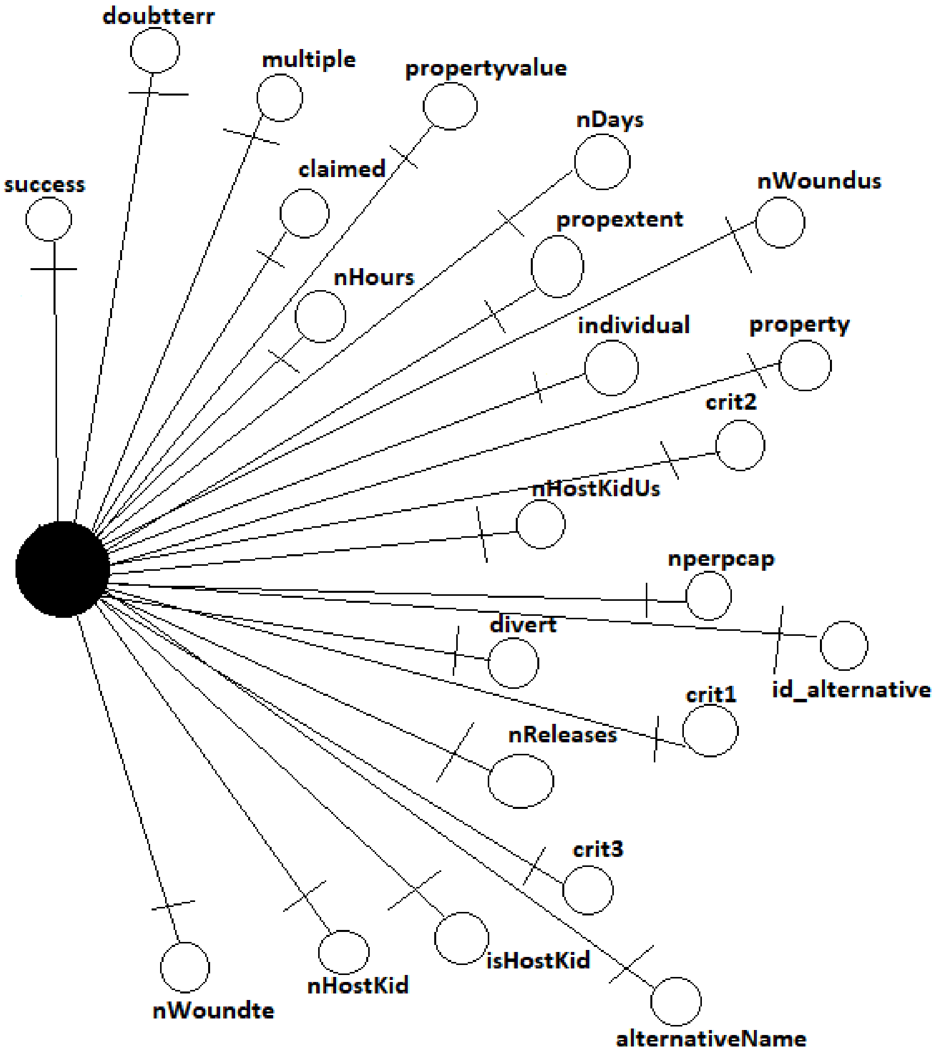


Figure 3: Tree cuts pt. 1

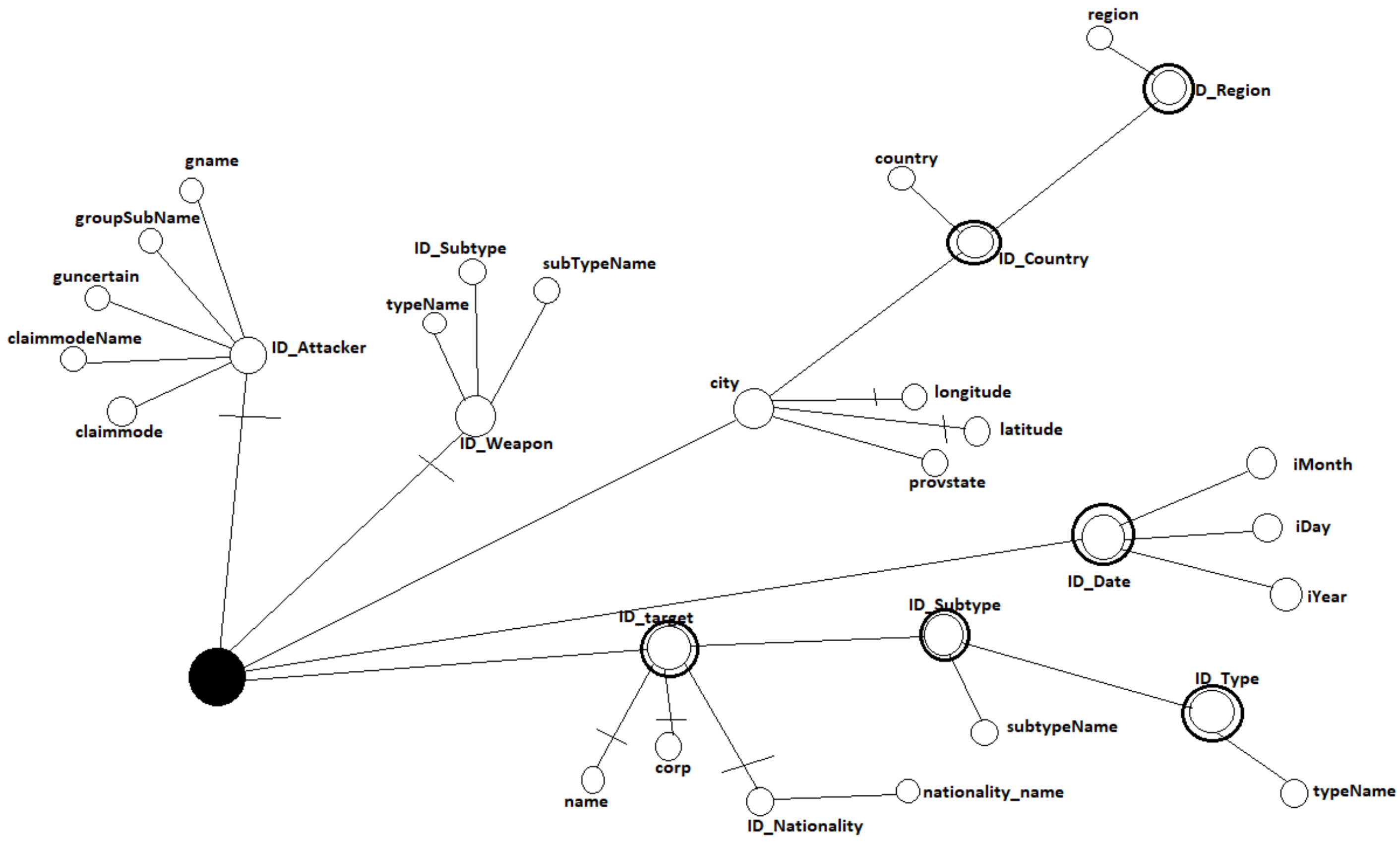


Figure 4: Tree cuts pt. 2

3.2 Final tree

Therefore the final tree is this:

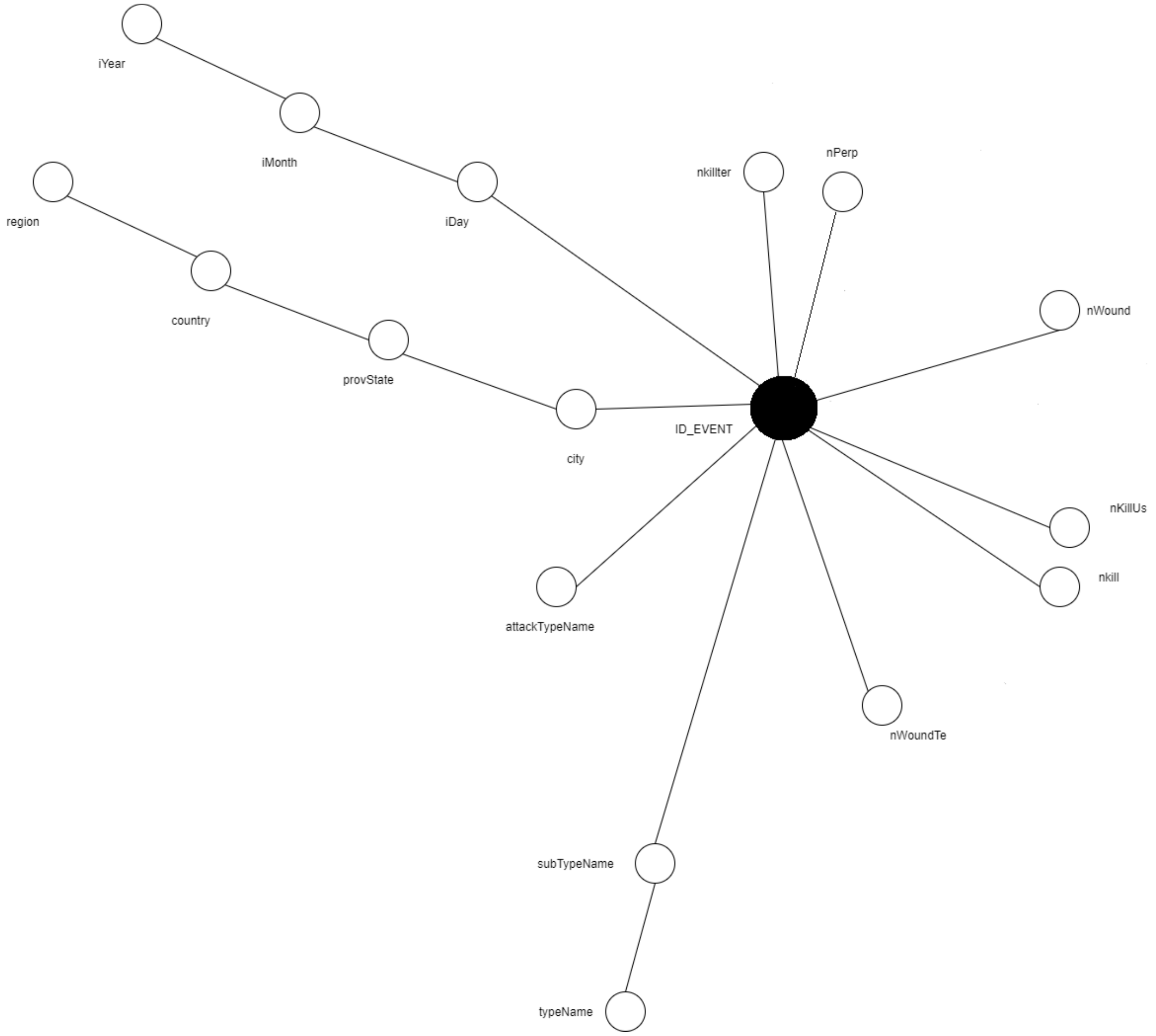


Figure 5: Final Tree

3.2.1 Dimensions and fact schema

The dimensions are:

- Date
- Target
- Place
- Attack type

The fact table is:

- Event

The final schema :

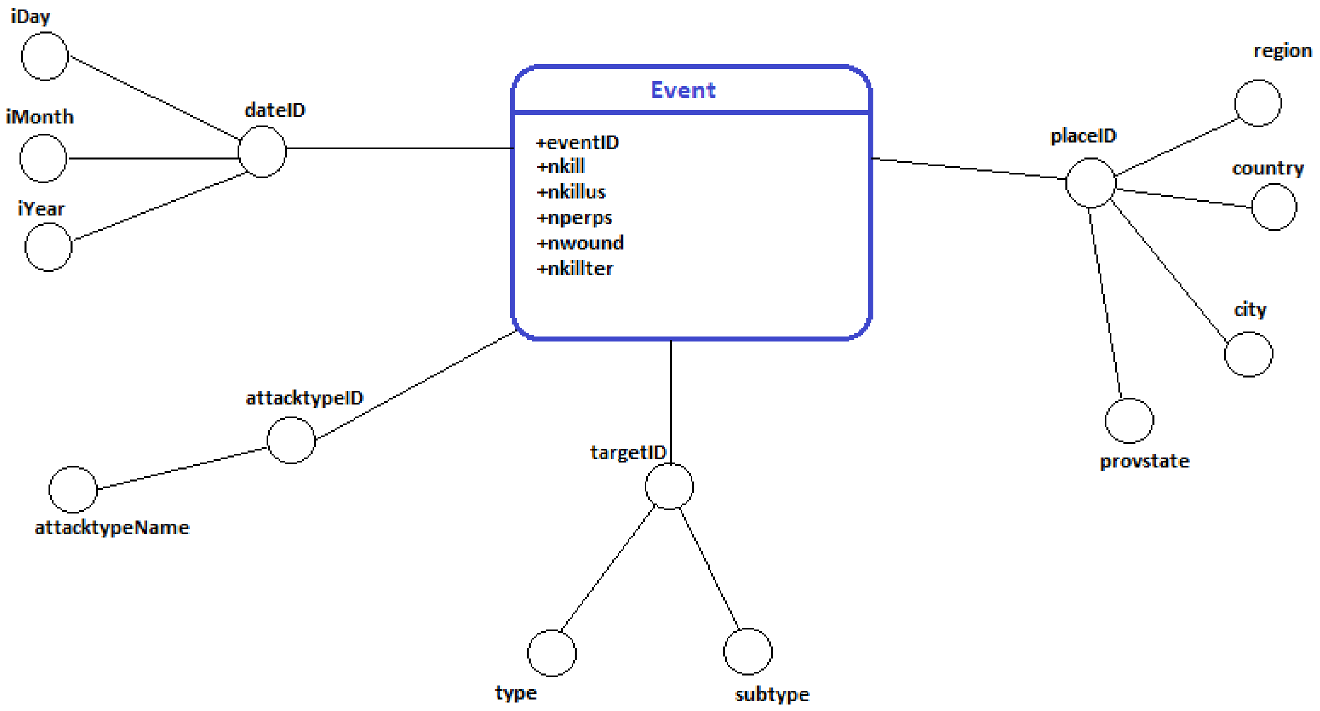


Figure 6: Final schema

3.3 Fact schema

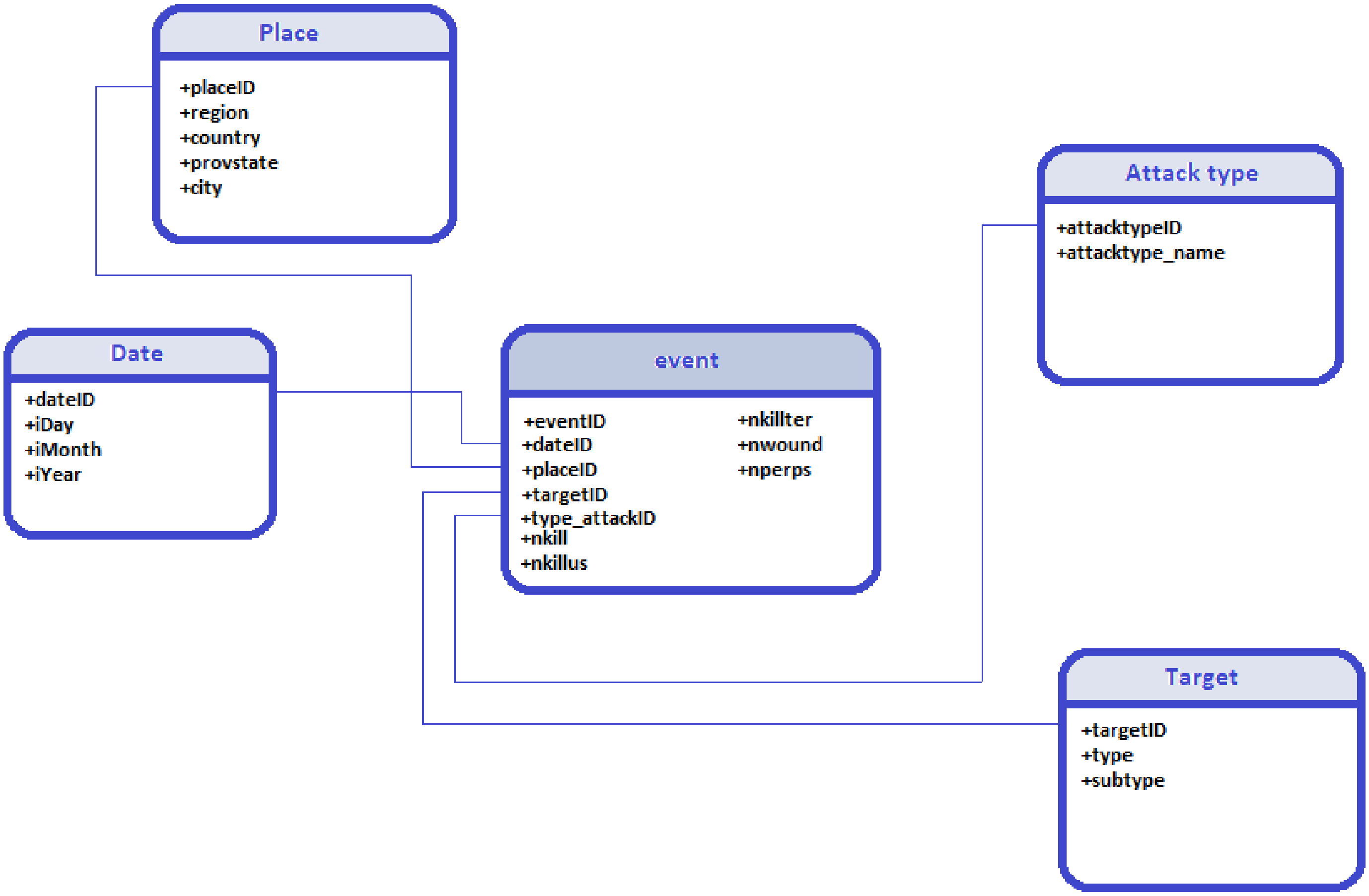


Figure 7: Fact schema

4 Data cleaning and preprocessing

Data preprocessing involves the transformation of the raw dataset into a clear and understandable format and it is fundamental to improve data efficiency as it directly affects the outcomes of the analysis. As a tool for this steps I used Pentaho Data Integration (PDI) Client (Spoon), which provides the Extract, Transform and Load (ETL) capabilities that facilitates the process of capturing, cleansing, and storing data.

4.1 Extraction, Cleaning and Transformation

The first step on Pentaho was to retrieve data from a csv file, and after cleansing and transforming the data, a new file with the new formatted information has been created. The exact steps in order that have been performed are the following:

1. Cleaning by keeping only the columns I'm interested in
2. Sorting the records by event ID
3. Removing double rows and leaving only unique occurrences
4. Mapping null values in the day field from 0 to 1
5. Mapping null values in the month field from 0 to 1

6. Filtering rows making sure that day, month, country, region, provstate and city are NOT NULL and that provstate and city are not "Unknown"

7. Mapping nperps (the number of terrorist which participated in the attack) from -99 or empty value to 0

8. Mapping the target type from empty value to "Unknown"

9. Mapping the target subtype from empty value to "Unknown"

10. Mapping the number of wounds from empty value to 0 and from 8.5 to 8, because I noticed there was this particularly strange floating point number

11. Mapping nkill (the number of total fatalities) from empty values to 0

12. Mapping nkillus (the number of total us fatalities) from empty values to 0

13. Mapping the number of terrorists that died in the attack from empty values to 0

14. Sorting the records by date in ascending way

15. I chained day month and year in a new field called "data"

16. Data has been reported in a new csv file

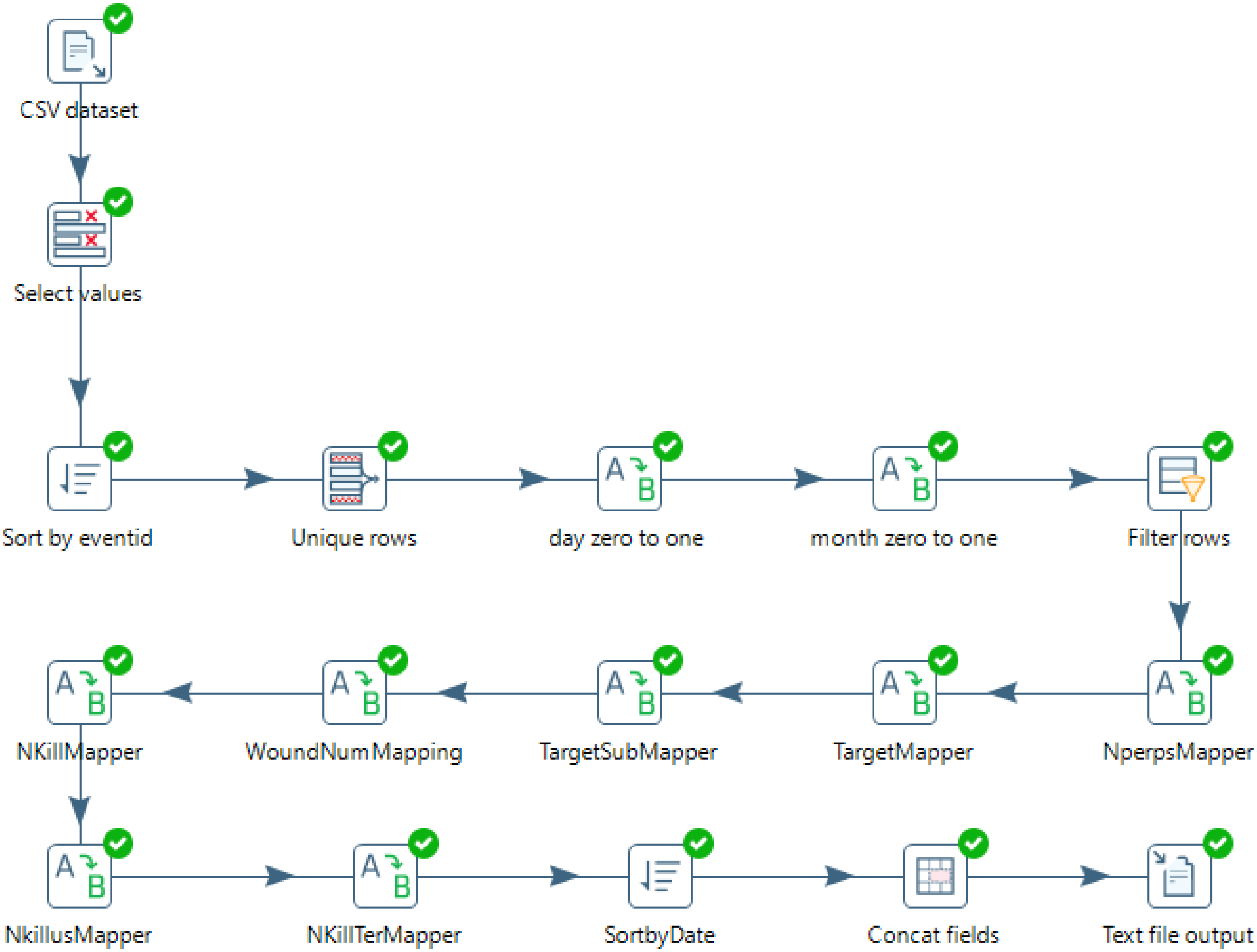


Figure 8: Cleaning and transformation

4.2 Load

Then I loaded the data about my dimensions: date, place, target and type of attack. The schema looks the same for each of them. This is the one for date:

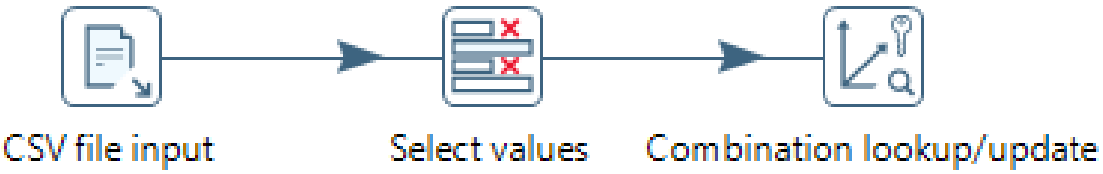


Figure 9: Date load

Each of these four schemes performs the following steps:

1. Takes as input file the one that has been generated by the cleaning and transforming file
2. Select the values to keep and those to discard (in this case, for the table ”data” it will only keep the field ”data” and discard the others)
3. Updates the dimension

At this point the only table left to create was my event table and the schema is this one:

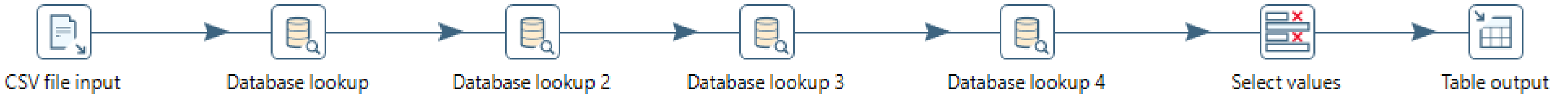


Figure 10: Event

What is performed here is:

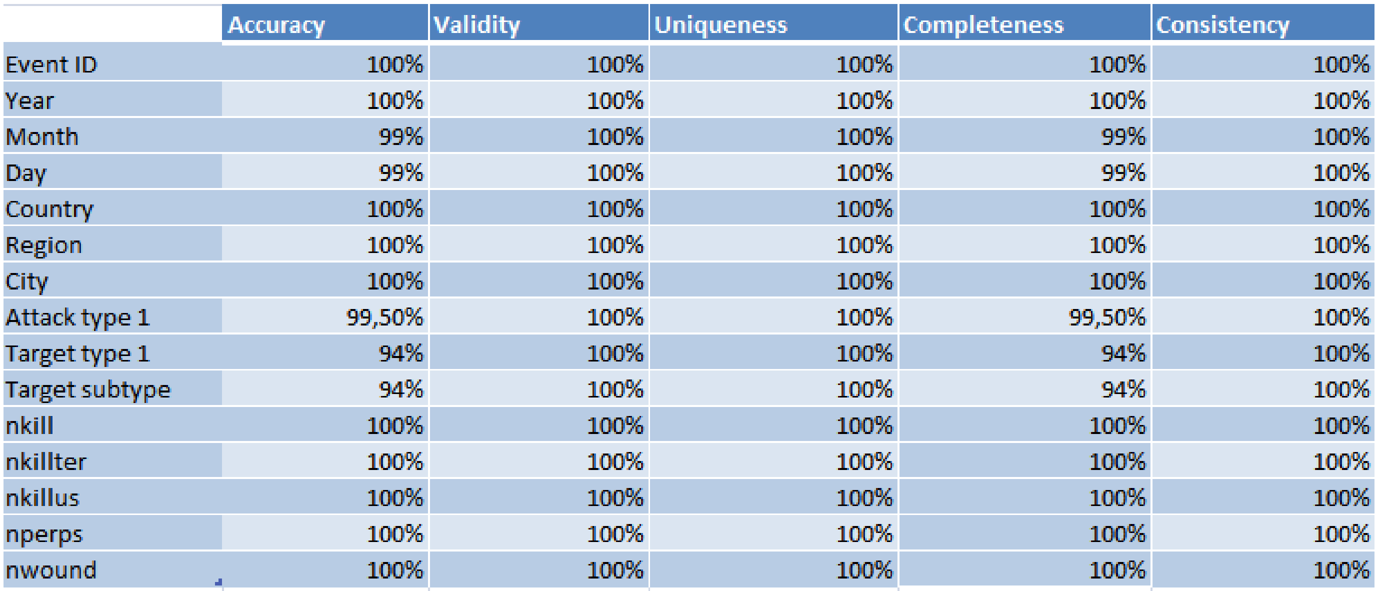
1. Look up date values in the database
2. Look up place values in the database
3. Look up target values in the database
4. Look up attack type values in the database
5. Removing unnecessary fields
6. Writing the information into a new table called ”event” in the database

4.3 Data quality

Data quality is the most insidious and important aspect to consider in order to create a solid analytic system and data is generally considered to be of high quality if it is suitable for its intended uses in operations. When we talk about data quality we are not talking about a single parameter, it is a factor that depends on many aspects, such as: missing values, wrong values, etc... The data quality metrics are:

- Accuracy: Data in the database systems matches agreed on facts about it in the world.
- Completeness: Data needed to describe entities in data sets exists in the database system.
- Consistency: when datasets move from one location to another, their contents remain the same.
- Validity: Data are valid if it conforms to the syntax (format, type, range) of its definition.
- Uniqueness: nothing will be recorded more than once based upon how that thing is identified.

After the ETL process I did a bit of analysis to see the level of data quality, and based on the result I understood if there was something to change / improve. Below the results table: As we can see from the table the values are very good, the only ones to have a quite notable defect are the target and the subtarget field as they have some lines with the value "Unknown". Furthermore, the day field is not 100 percent accurate as in the transformation phase, as there are some values set to zero, they have been changed with one, therefore the attacks with day one is not totally reliable that they really took place on the first day of that month (the same for month field).



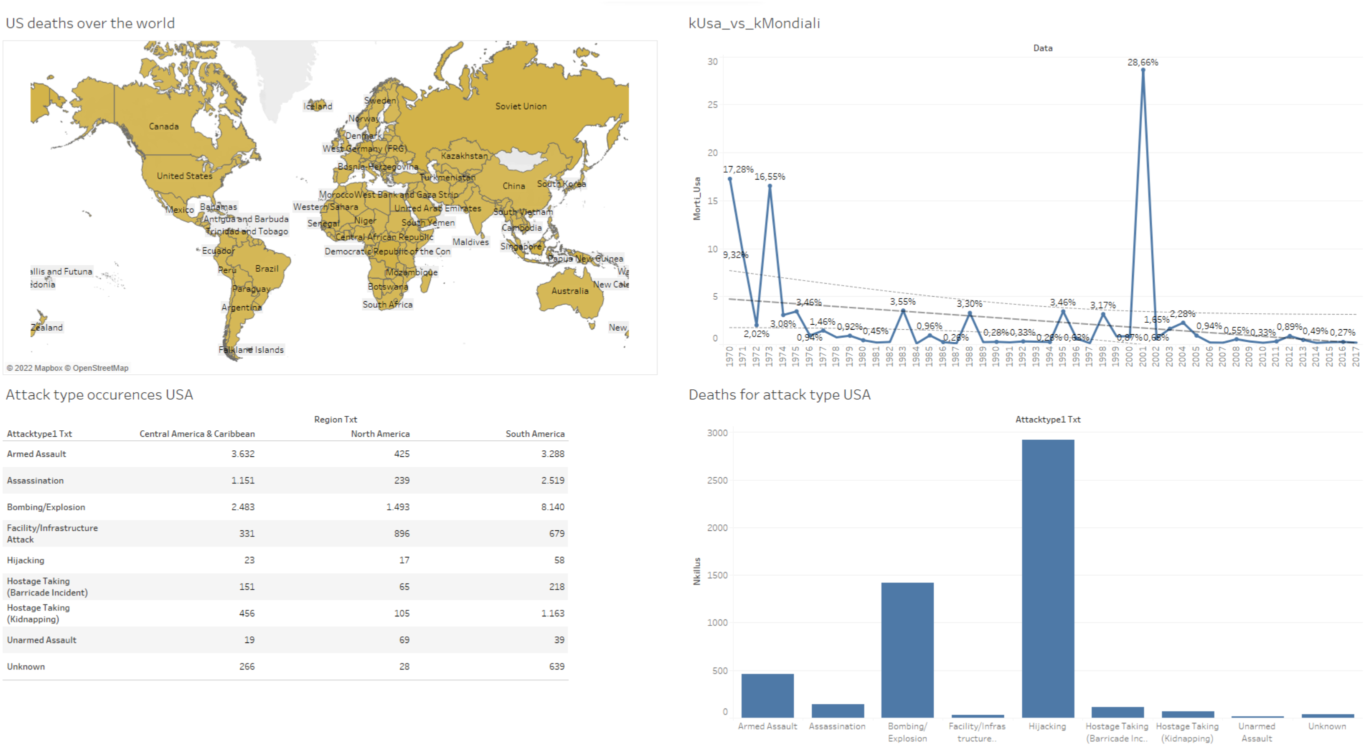
	Accuracy	Validity	Uniqueness	Completeness	Consistency
Event ID	100%	100%	100%	100%	100%
Year	100%	100%	100%	100%	100%
Month	99%	100%	100%	99%	100%
Day	99%	100%	100%	99%	100%
Country	100%	100%	100%	100%	100%
Region	100%	100%	100%	100%	100%
City	100%	100%	100%	100%	100%
Attack type 1	99,50%	100%	100%	99,50%	100%
Target type 1	94%	100%	100%	94%	100%
Target subtype	94%	100%	100%	94%	100%
nkill	100%	100%	100%	100%	100%
nkillter	100%	100%	100%	100%	100%
nkillus	100%	100%	100%	100%	100%
nperps	100%	100%	100%	100%	100%
nwound	100%	100%	100%	100%	100%

Figure 11: Data quality

5 Analysis sheets and Dashboards

To build the analysis sheets and the dashboards I used Tableau, which is a powerful and fast tool for data visualization. The analysis I made are the following:

5.1 USA dashboard



Attack type occurences USA

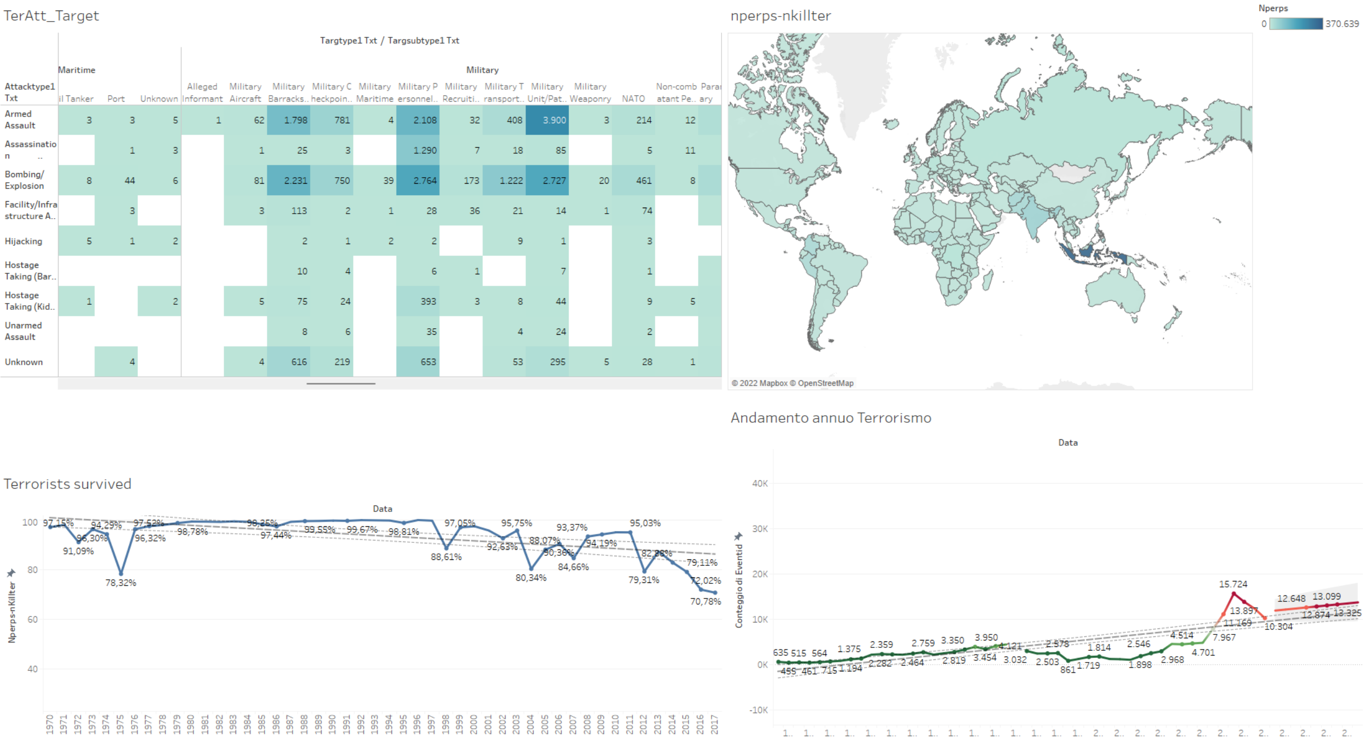
		Region Txt	
Attacktype1 Txt	Central America & Caribbean	North America	South America
Armed Assault	3.632	425	3.288
Assassination	1.151	239	2.519
Bombing/Explosion	2.483	1.493	8.140
Facility/Infrastructure Attack	331	896	679
Hijacking	23	17	58
Hostage Taking (Barricade Incident)	151	65	218
Hostage Taking (Kidnapping)	456	105	1.163
Unarmed Assault	19	69	39
Unknown	266	28	639

Figure 12: Dashboard USA

In this first dashboard I put together the sheet (on the upper left of the picture) where there is the world map and by hovering over any state with the cursor appears a label that tells the name of the state and the number of deaths, while the sheet on the upper right shows the percentage per year of American deaths in relation to world deaths, and the one on the lower right shows the frequency of attack type in the USA. From this dashboard we can observe that, obviously, the peak was in 2001 and that hijacking was the type of attack that resulted in the most deaths despite the fact that the bombing/explosion is the most frequent type of attack in the USA counting over 11.000 bombings against less than 100 hijackings. Furthermore we can observe that in the last years the percentage of deaths has decreased a lot.

5.2 Terrorists dashboard

This second dashboard is also composed of four sheets of which the first in the upper left shows the occurrences for every combination of attack type and the target of it, while the lower left one shows in percentage how many terrorists survived every year above all those who participated in the attacks. The upper right sheet shows a world map where if you hover on any state with the cursor appears a label that tells the name of the state, the total number of terrorists that



TerAtt_Target

Targtype1 Txt / Targsubtype1 Txt

Attacktype1 Txt	Maritime			Military												
	il Tanker	Port	Unknown	Alleged Informant	Military Aircraft	Military Barracks..	Military Checkpoin..	Military Maritime	Military Personnel..	Military Recruiti..	Military Transport..	Military Unit/Pat..	Military Weaponry	NATO	Non-combatant Pe..	Para ary
Armed Assault	3	3	5	1	62	1.798	781	4	2.108	32	408	3.900	3	214	12	
Assassination ..		1	3		1	25	3		1.290	7	18	85		5	11	
Bombing/ Explosion	8	44	6		81	2.231	750	39	2.764	173	1.222	2.727	20	461	8	
Facility/Infrastructure A..		3			3	113	2	1	28	36	21	14	1	74		
Hijacking	5	1	2			2	1	2	2		9	1		3		
Hostage Taking (Bar..						10	4		6	1		7		1		
Hostage Taking (Kid..	1		2		5	75	24		393	3	8	44		9	5	
Unarmed Assault						8	6		35		4	24		2		
Unknown		4			4	616	219		653		53	295	5	28	1	

nperps-nkillter

© 2022 Mapbox © OpenStreetMap

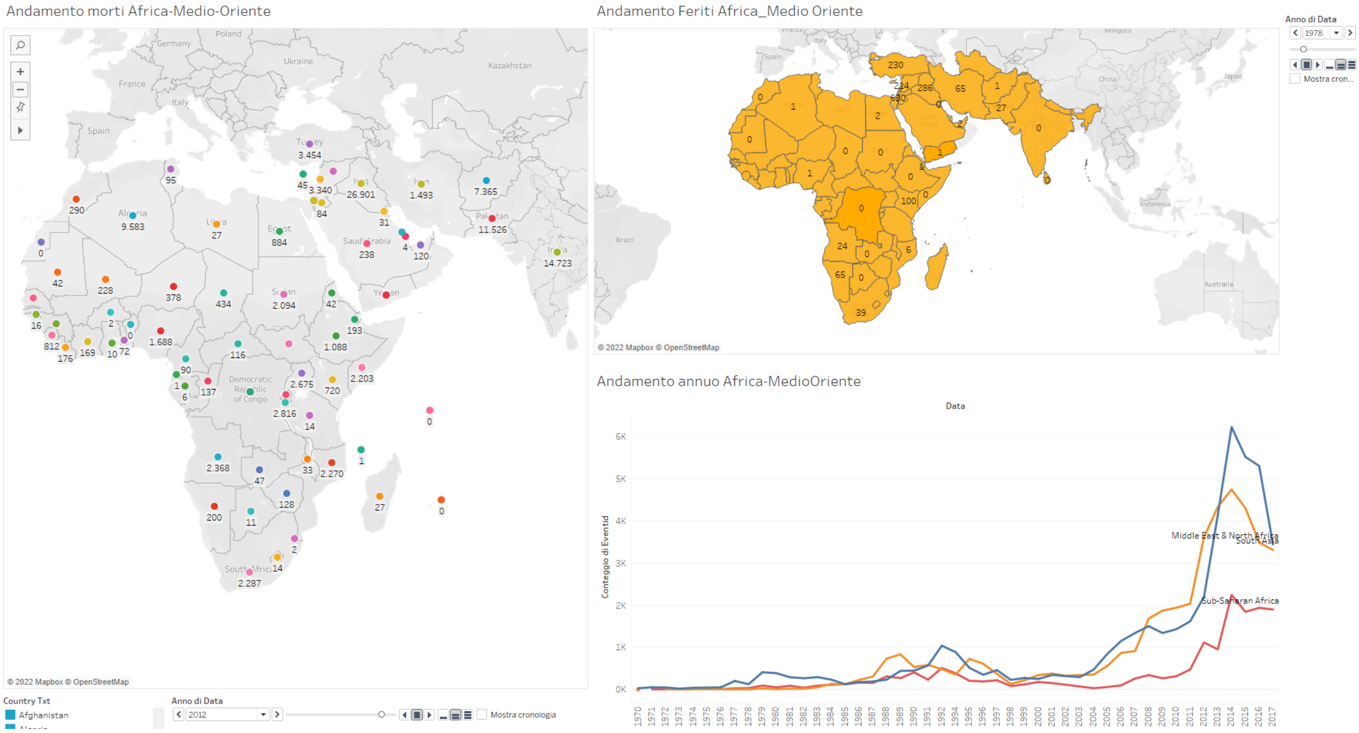
Terrorists survived

Andamento annuo Terrorismo

Figure 13: Dashboard terrorists

participated in the attacks and the number of them who died. The last sheet, the lower right one, shows the total number of terrorist attacks every year until 2017 and then makes a rough estimate of possible future attacks from 2018 to 2025. Obviously this is an estimate that the software makes only based on the trend of the values of previous years, but out of curiosity I still searched on the internet for the number of terrorist attacks in 2018 and it was 15,321 against the 11,970 expected. From this dashboard we realize that bombing/explosion is the most frequent attack type, especially with target military and police, and also we note that usually 90 percent of terrorists always survive.

5.3 Middle East and Africa dashboard



Andamento morti Africa-Medio-Oriente

Andamento Feriti Africa_Medio Oriente

Andamento annuo Africa-MedioOriente

Figure 14: Dashboard middle East and Africa

This third dashboard is about middle East and Africa, in fact in the left sheet shows with an animation for every state the number of deaths for every year incrementally, while the upper right sheet does the exact same thing but for the number of injured. Finally, the lower right sheet shows for middle east and north Africa, south Asia and Sub-Saharan Africa the number of attacks by year and from this one we can see how it really increased in the last years.

5.4 Deaths and injured dashboard

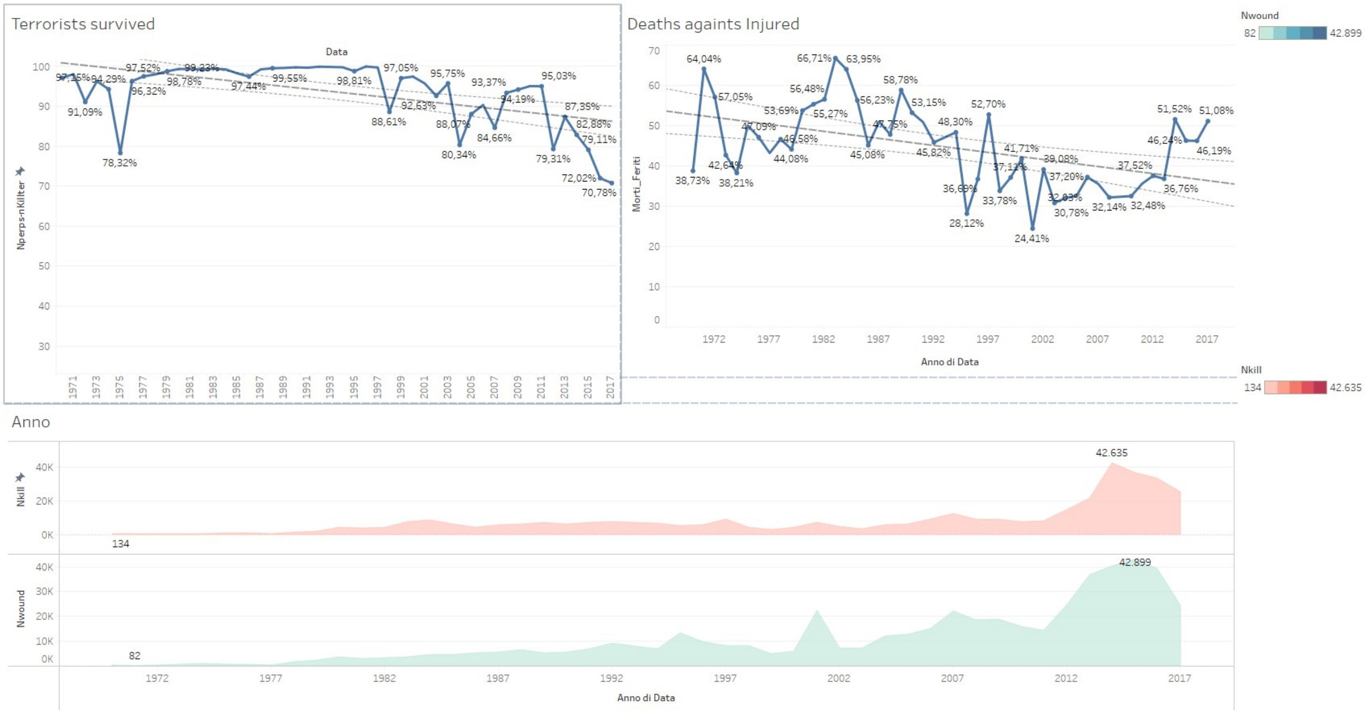


Figure 15: Dashboard deaths

This dashboard shows in the sheet on the upper left the percentage of survived terrorists per year, and as we can see the percentage never drops below 70 percent, while in the upper right sheet we can see the percentage of deaths against the injured people, always by year. The sheet on the bottom shows on the upper row the trend of deaths, while the lower one the injured trend, always by year.

6 Conclusions

The National Consortium for the Study of Terrorism and Responses to Terrorism (START) is continuing to track global attacks each year, and from what can be found on the internet from their articles on the latest years the data shows that deaths from terrorist attacks are decreasing every year, and that unfortunately Afghanistan continues to be in first place among the countries with the most terrorism. It is also interesting to observe that since the outbreak of the pandemic, terrorism has been changing: the Counter-Terrorism Committee Executive Directorate's (CTED) 2021 report concludes noting that the pandemic has exacerbated many pre-existing issues and challenges that shape the terrorist threat landscape. Terrorists and violent extremists have sought to exploit pandemic-related socio-cultural restrictions, including their efforts to recruit, radicalize, and organize via virtual platforms.