YZV 475E - Term Project Report

Doruk Kurt, 150200325
Mert Gülşen, 150200332
Ege Demir, 150200319

04/01/2024

1 Dataset

Dataset for this project is a global terrorism database. This open-source dataset contains detailed information on terrorism events on a global scale from 1970 to 2017. It has various information like perpetrators, targets, tactics, outcomes, etc. There are over 100 variables and more than 180,000 attacks recorded. The dataset has date and coordinate information for each event therefore, the dataset is suitable for temporal and geospatial analysis.

Moreover, there are a lot of NaN values in the dataset. The image below shows the 10 columns that contain most NaN values and their NaN percentage.

```
Nan percentage:
gsubname3           99.988992
weapsubtype4_txt    99.961473
weapsubtype4        99.961473
weaptype4           99.959822
weaptype4_txt       99.959822
claimmode3          99.926799
claimmode3_txt      99.926799
gsubname2           99.911938
claim3              99.824978
guncertain3         99.823877
```

1.1 Key Variables

As we said, the dataset contains more than 100 columns, and we decided on some key variables between them to work on. Event date, event coordinates, country, number of casualties, target type, and attack type were among the key variables we picked. We will make our analysis mostly on them in the next parts.

2 Exploratory Data Analysis

2.1 Data Cleaning and Transformation

As we said when we mentioned the dataset, some columns have a lot of NaN values.

Firstly, we removed the columns that contain NaN values in more than half of their records to clean the data. Only 58 out of 135 columns remained.

Secondly, we created another column called 'total casualties' by simply summing the number of wounded victims and number of killed victims. We used this column in most of our plots.

Lastly, we created the date column by parsing eventid column that contains the date information. We used this feature too in most of our plots. These operations can be seen in the image below:

Data Cleaning

```
[ ]: # Removing columns where more than half of the values are NaN
half_len = len(df) / 2
df_reduced = df.dropna(thresh=half_len, axis=1)

# Displaying the shape of the original and reduced dataframes
original_shape = df.shape
reduced_shape = df_reduced.shape

original_shape, reduced_shape
```

```
[ ]: ((181691, 135), (181691, 58))
```

Data Transformation

```
[ ]: df_reduced["ncasualties"] = df_reduced["nwound"] + df_reduced["nkill"]
```

```
[ ]: def parse(x):
    x = str(x)
    year = x[0:4]
    month = x[4:6]
    day = x[6:8]
    return (f"{day}/{month}/{year}")
df_reduced["date"] = df_reduced["eventid"].apply(parse)
```

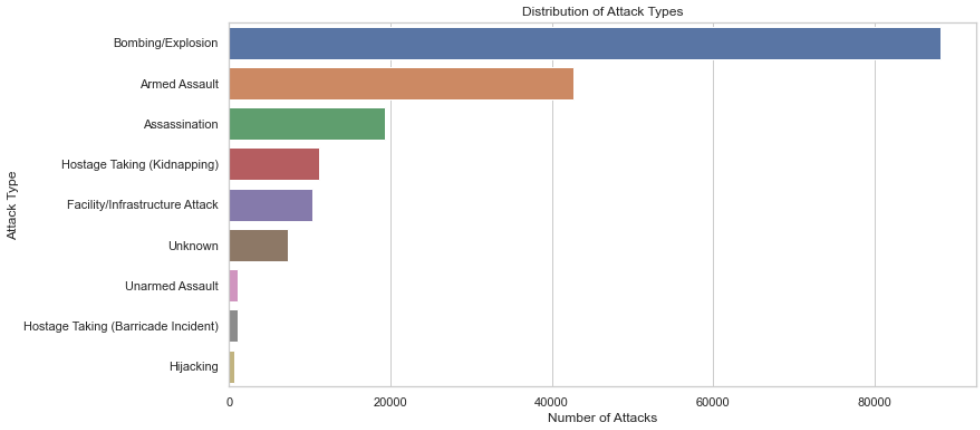
```
[ ]: # Specify the path to the new CSV file
new_csv_path = 'cleaned_data.csv'

# Save the first 5000 rows to the new CSV file
df_reduced.to_csv(new_csv_path, index=False)
```

After completing the steps, new dataframe is saved to a new csv file. This file "cleaned_data.csv" is then used as data source in tableau workbook.

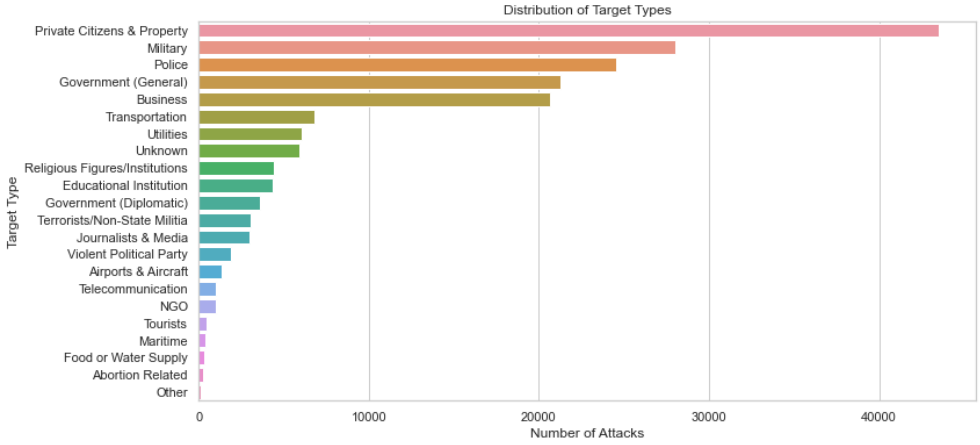
2.2 Distribution of Key Variables

Let's see the distribution of the key variables.

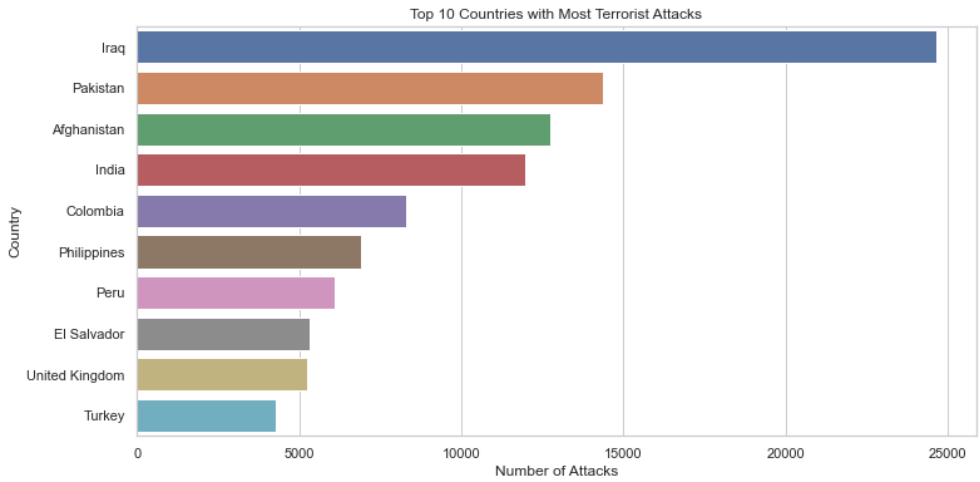
As seen below, bombing is the most popular attack type among terrorists, followed by armed assault and assassination attacks.



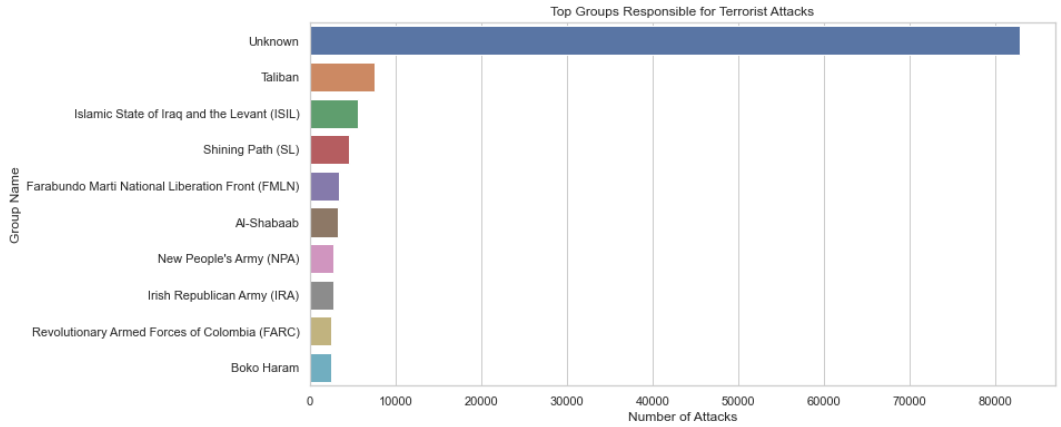
Terrorists mostly target private citizens. Military, police, and government are also among mostly targeted groups.



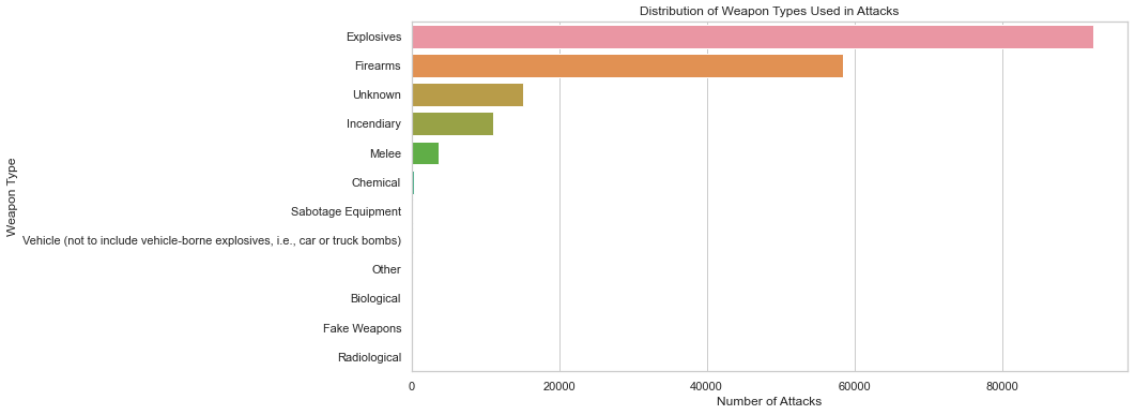
Iraq, Pakistan, and Afghanistan are the top 3 countries in terms of number of terror attacks. Sadly, Turkey is also in top 10.



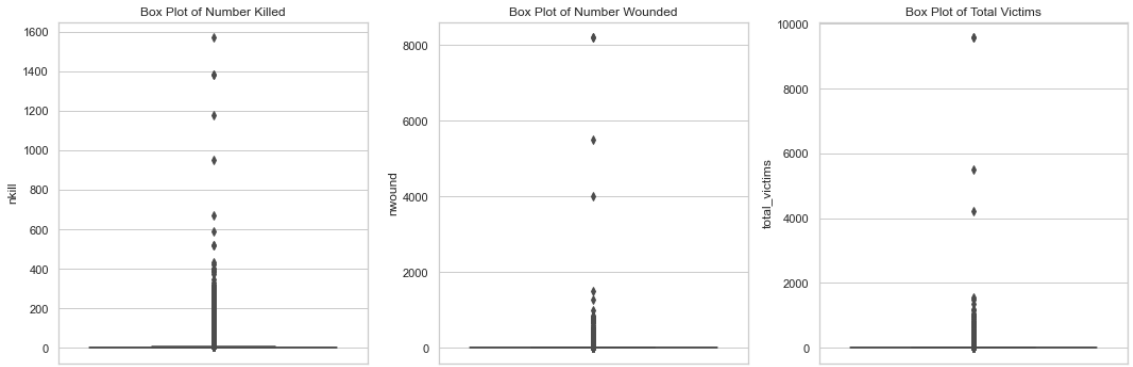
For most of the events, the responsible terror group is unknown. Among the attacks that the responsible group is known; Taliban, ISIL and Shining Path terror groups are the most active groups.



Explosives and firearms are the most popular weapons by far, followed by incendiary and melee weapons.

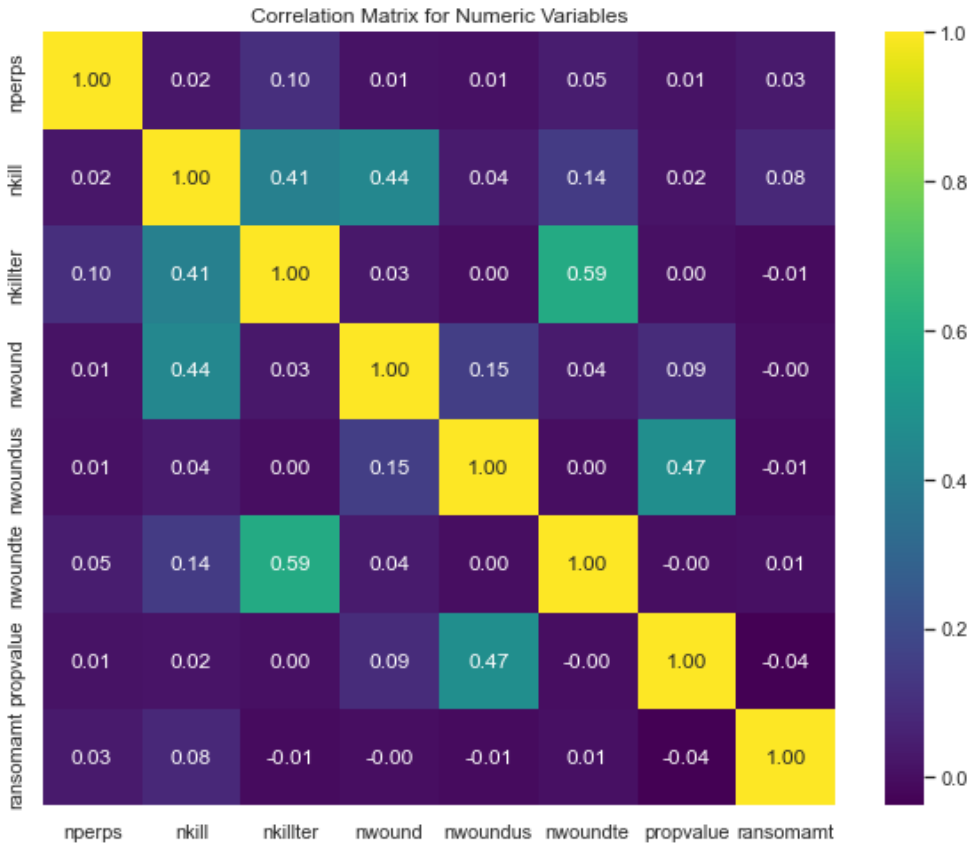


Lastly, let's check the box plot of the number of killed, the number of wounded, and the number of victims columns. Most of the data points(each of them visualizes a specific attack), reside at the bottom of the plot. The most destructive attacks, in terms of victim numbers, are the outliers of these plots. For example 9/11 attack is at the highest for 2 of these plots and it's at the 2nd place for the other.



2.3 Correlation Matrix

Even though most of our columns are categorical, we have some numeric features too. Let's check the correlation between them:



This correlation matrix tells us that there is a significant correlation between the number of killed terrorists and the number of wounded terrorists. Similarly, there is a significant correlation between the number of killed victims and the number of wounded victims. Both of these correlation makes sense.

In a more surprising correlation, we see that the number of wounded US citizens is positively correlated with the cost of attack in US dollars. It seems that when US citizens get wounded, it damages the economy more.

2.4 EDA Findings

Let's sum up the findings we get from this analysis.

1. Bombings using explosives are the most common attacks.
2. A lot of the casualties happen in a few of the attacks. We can call these events 'outliers'.
3. When the number of people killed is high, number of injured people is likely to be also high.
4. We don't know the responsible groups for most of the terror events.
5. Middle Eastern and South Asian countries are in great danger in terms of the number of terror events.

6. Even though private citizens are the targets of most of the attacks, soldiers and police are in greater danger, proportionally speaking.

3 Tableau Plots

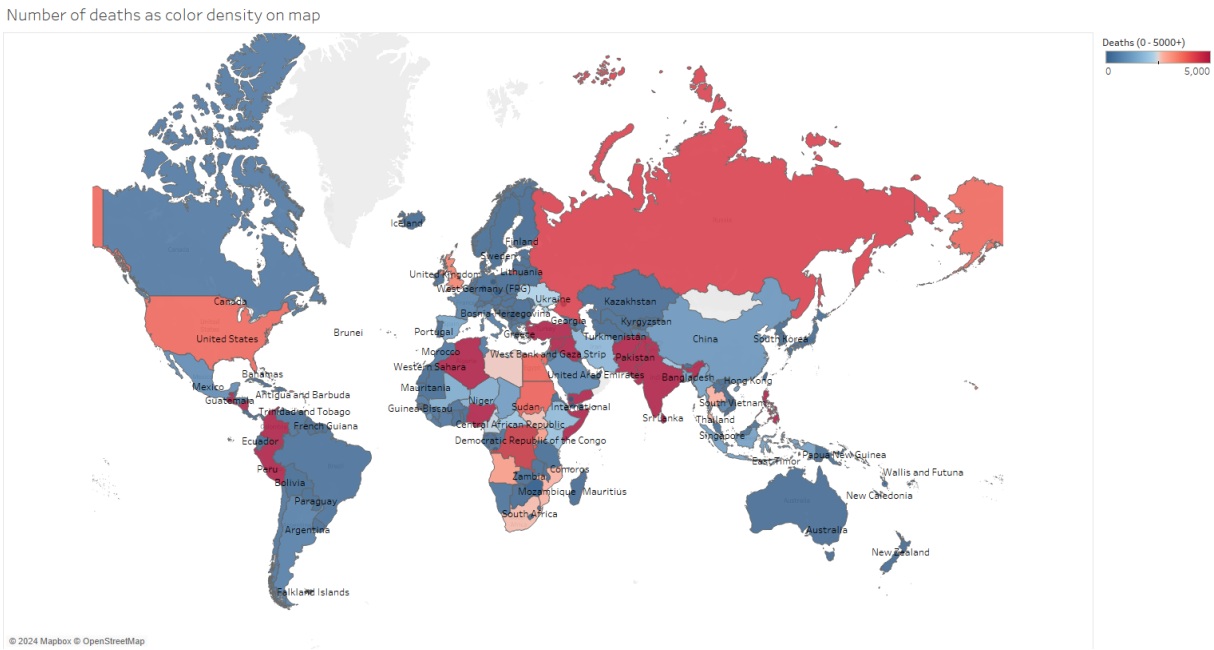
3.1 Introduction

In this section, plots that are created using Tableau software are explained. Some plots are displayed on a dashboard while some plots are directly displayed as a worksheet. All of the plots are placed on a Tableau Story and presented. Note that in some plots, null values are excluded using Tableau's exclude option.

3.2 Number of deaths as color density on map

In the figure below, a geospatial plot is displayed. This plot shows the world map with color densities. For each country, color densities represent the total number of deaths caused by terror events. Darker red color implies more deaths and darker blue implies less deaths.

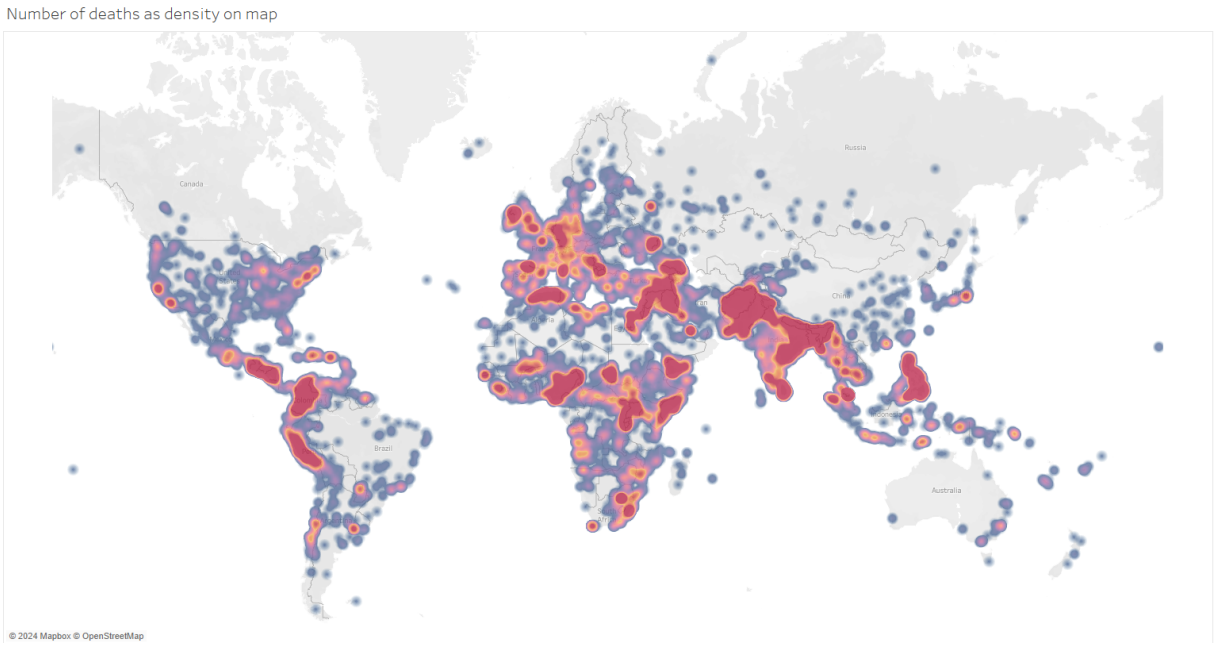
For this plot and other geospatial plots, each country on the world map is represented using the average longitude and latitude columns in the dataset. To group and label terror events for every country, country_txt column is used which contains name of the country for that terror event. Finally, to create the color density, sum aggregated nkill (number of deaths) column is used as color indicator from Marks section in Tableau.



3.3 Number of deaths as density on map

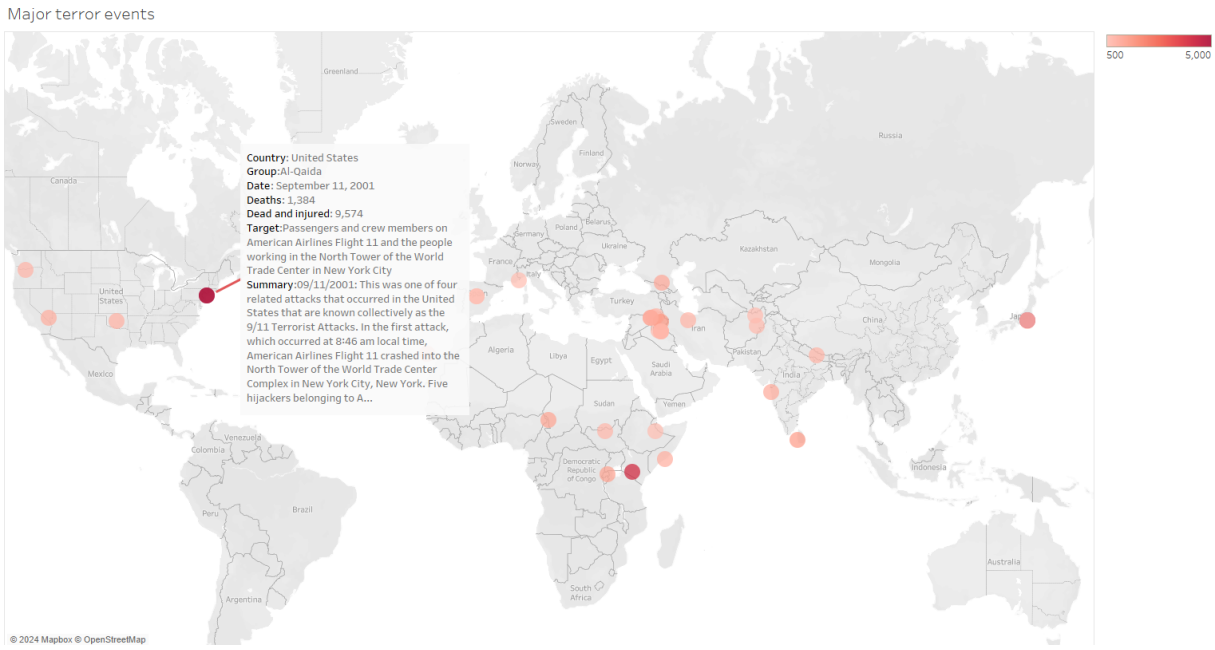
In the plot below, number of terrorist attacks are displayed on a world map as densities. More densely colored area means more terrorist attacks in that area. Similar to the first geospatial

plot, a world map is created with average longitude and latitude columns. In this plot, count aggregated eventid column is used and grouped by country names. This ensures each country represent the count of terror events in that country.



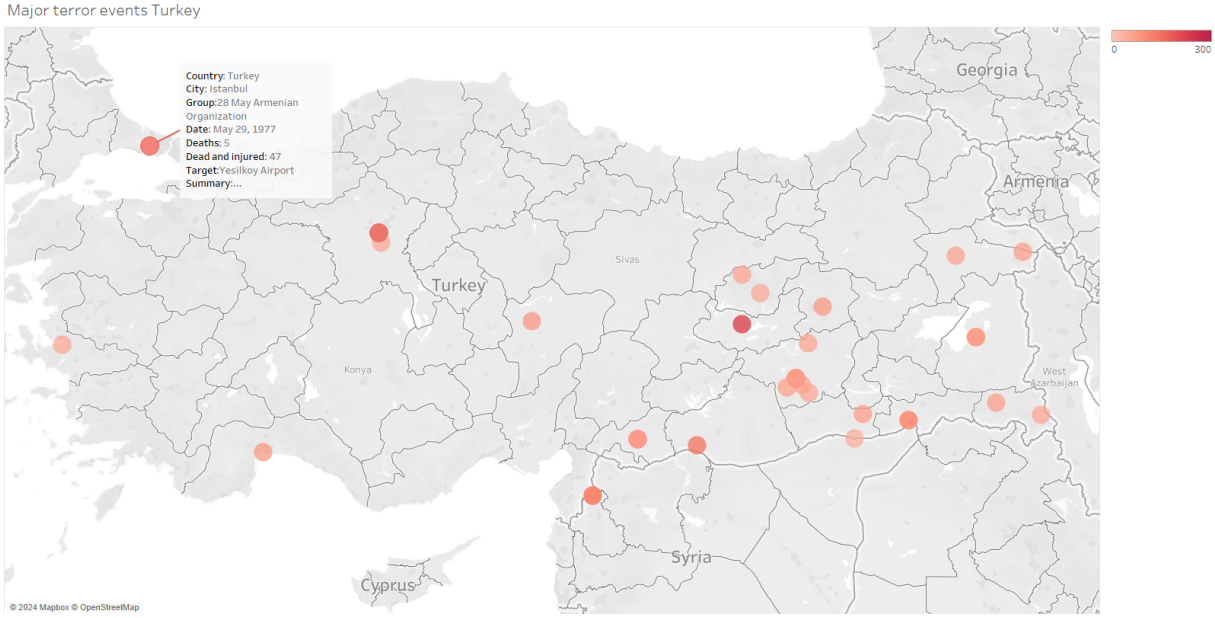
3.4 Major terror events

In the plot below, terror events that have caused the most casualties are displayed as circles on a world map. Finding these events are done by using the filter section in tableau. Previously calculated "total casualties" column is used as filter and events that have casualties between 500 (threshold) and 9574(max) are displayed on the world map. Circles are also colored with number of casualties where darker colored events represent more casualties. To display information of an event when mouse hovers over the circle, tooltip is configured with information related to events like country name, terrorist group, date etc.



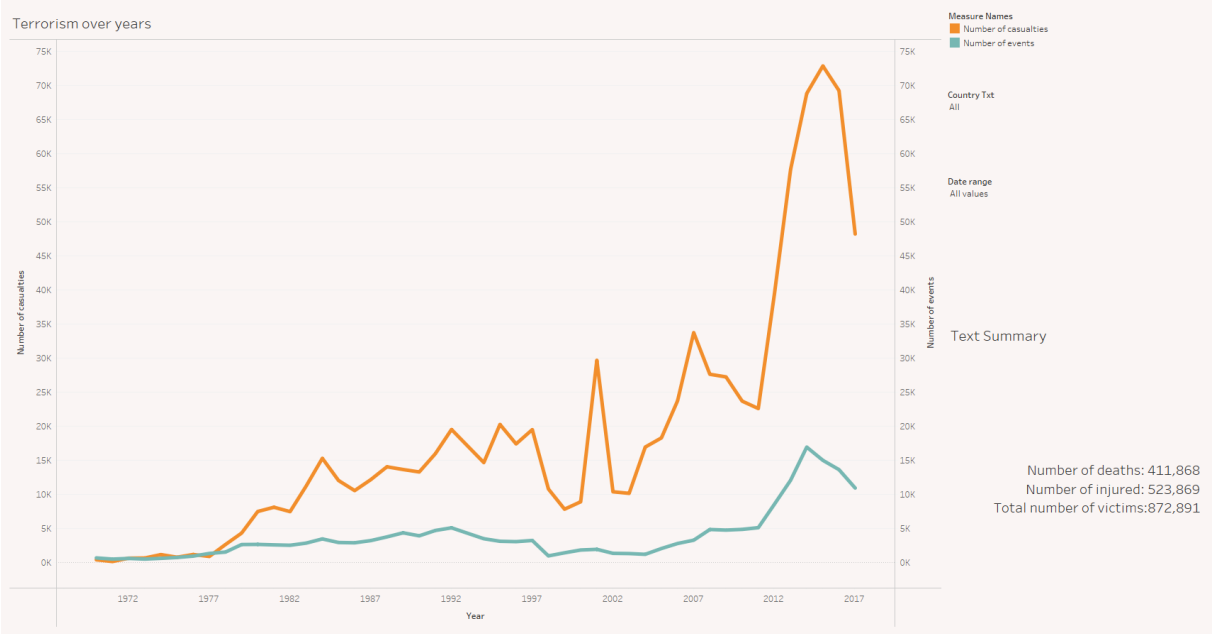
3.5 Major terror events in Turkey

In the next plot below, major terror events in Türkiye are displayed similar to the previous plot. Using the same methods and adjusting the number of casualties filter, terror events that caused the most casualties are displayed as circles on the map of Türkiye. Tooltip of data points are configured to display information about terror attacks and circles are density colored with number of casualties.



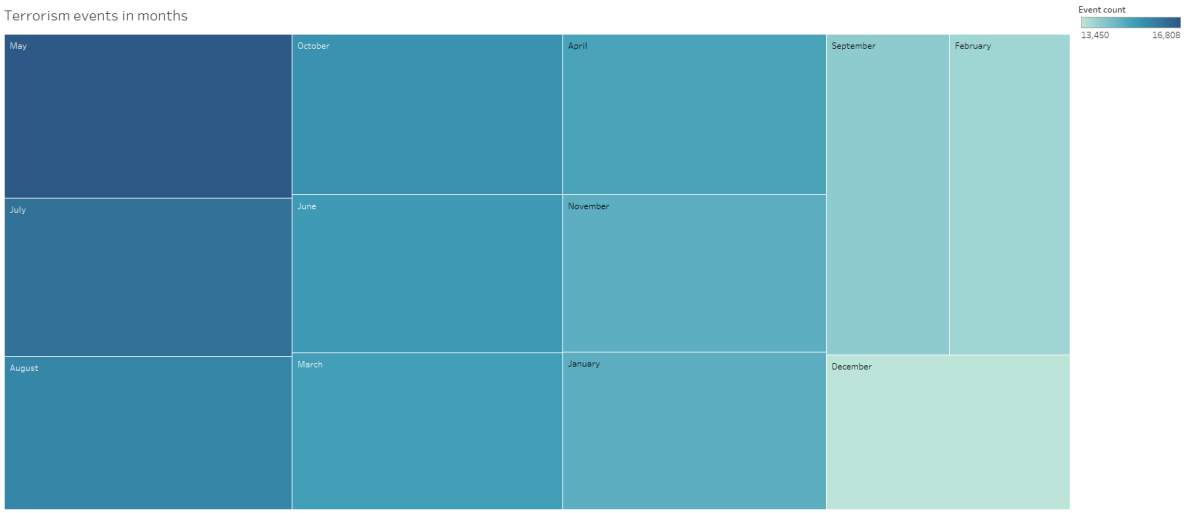
3.6 Terrorism over years

Next, a time series plot is created as a line plot with dual axis. In this plot, x axis shows the individual years, y axis to the left shows number of casualties and y axis to the right shows number of events in that year. From this plot, it can be concluded that there is an increasing trend in terrorism over the years. A special case in this plot is around year 2001 where 9/11 attack was done in USA. In the number of events line there are no apparent anomaly however in the casualty line, there is a sudden peak in number of casualties in that year. This indicates a terror event that has caused mass casualties. The plot can be modified with using the created country and date filters. Specified values can be selected to display the plot using filtered data.



3.7 Terrorism events in months

In the figure below, a treemap chart is shown to visualize terrorism events over specific months. Each rectangle in the chart represents a map. The size and color density of the rectangles represent the total number of terror events that happened in that month for all years. It can be seen that most terrorist acts happened in May and the least amount of terror events happened in December. It is also seen that in general, terror events happened most in months of the summer season and least in months of the winter season.



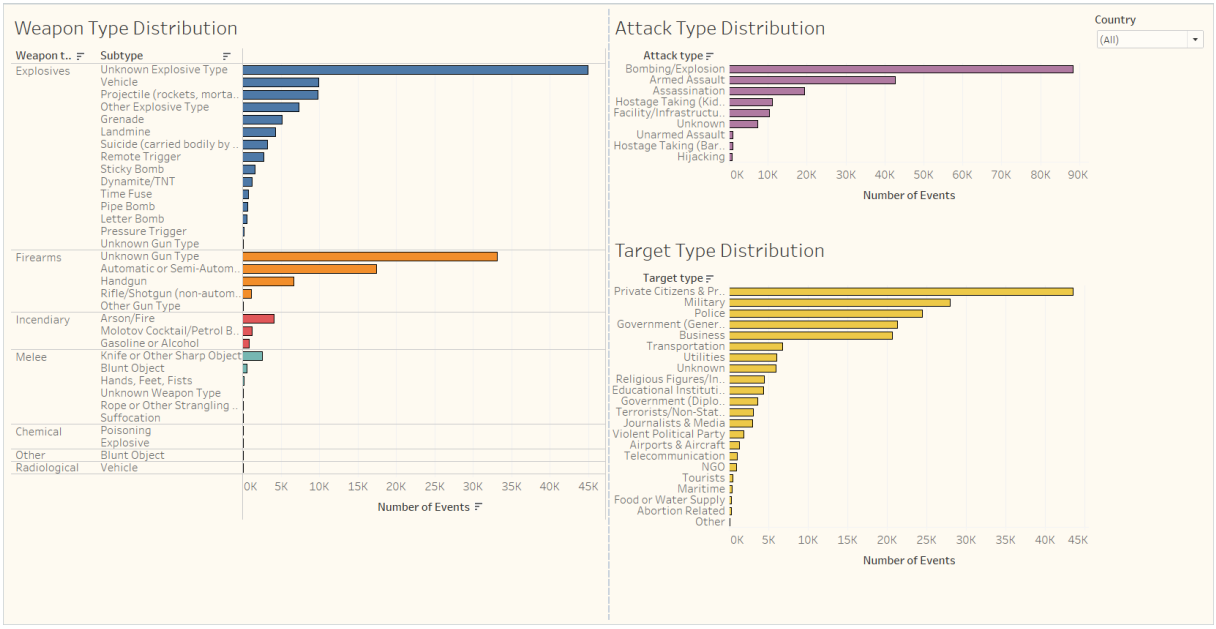
3.8 Bar plots for visualizing columns

The bar plots below display the distributions of the weapons used, types of terrorist attacks, and the targets of the attacks. In default, the bar plots show the distribution according to all countries combined. The country can be specified using the dropdown menu in the top right corner.

The bar plot on the left side is the distribution of the weapon types that are used in terror events. Each general weapon type has a different color. There are also weapon subtypes that give more specific insights about the type of the weapon. According to the weapon type distribution, the most used weapons are the explosive weapons. More specifically the most used weapons are unknown explosives, which indicates that most of the time the explosives cannot be fully determined by the authorities. The firearms come after the explosives which are also frequently used in terror events.

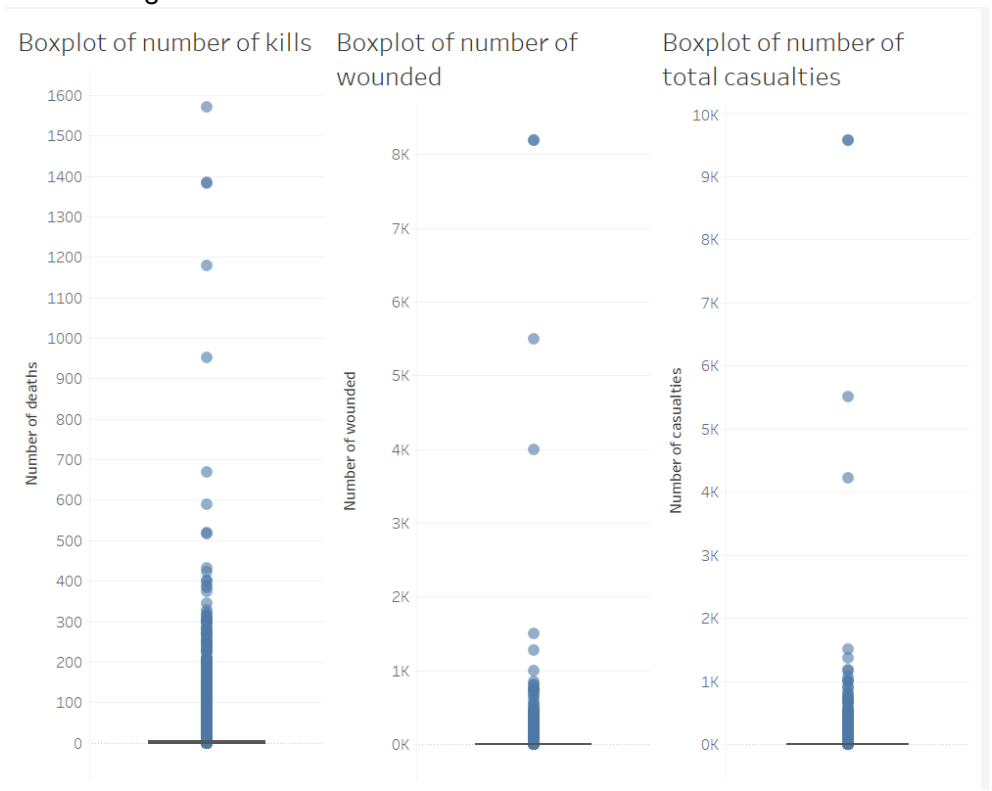
The top right bar plot displays the distribution of the attack types. The plot indicates that most of the terror attacks are in the form of a bombing or explosion. Armed assaults come in second place with nearly half as frequent as explosions.

Lastly, the plot in the below right shows the frequency of the terror attack targets. Sadly, the most frequent targets of terror attacks in the world are private citizens, the military, and the police.



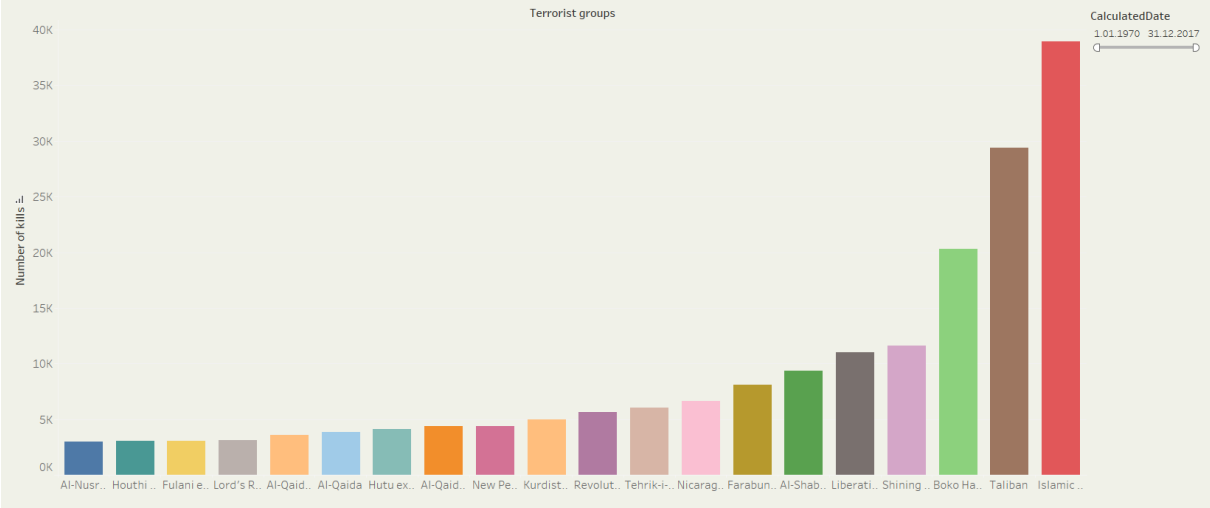
3.9 Boxplot analysis of key numerical columns

In the figure below three box plots can be seen. These box plots are created to analyze which terror events are outliers in terms of the number of deaths, the number of wounded, and the number of casualties. The y-axes are the number of deaths, number of wounded, and number of casualties respectively. Each point in the box plots represents one terror event. Sadly, from the plots, we can see that the 9/11 attacks and the attacks of ISIS in 2014 are the outliers in all boxplots due to the high number of casualties.



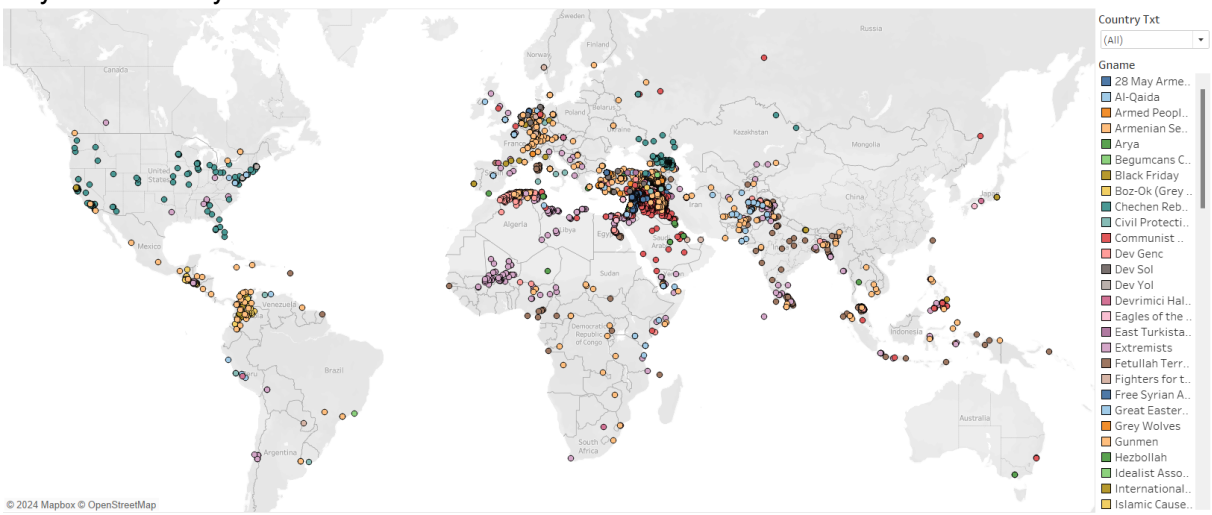
3.10 Bar plot analysis of terror groups

The bar plot below shows the terror groups with the highest number of kills in a specific period. The y-axis is the number of kills of the terrorist group in the specified time period. The terrorist groups are on the x-axis and different colors indicate different terror groups. The time period can be specified from the panel on the top left side which affects the groups shown in the plot. The bar on the rightmost side of the plot is the terrorist group that caused the most deaths in the given time period.



3.11 Terror groups and events displayed on a world map

The next figure, the one below is a geospatial plot that shows each terrorist event in the world with a dot. The dots show the location of the terrorist attack and each color of the dots represents the terrorist group. Initially, the world map is shown and all the events in the world are displayed. The country filter can be used to specify the country. In this way, the terror events in a specific country can be analyzed.

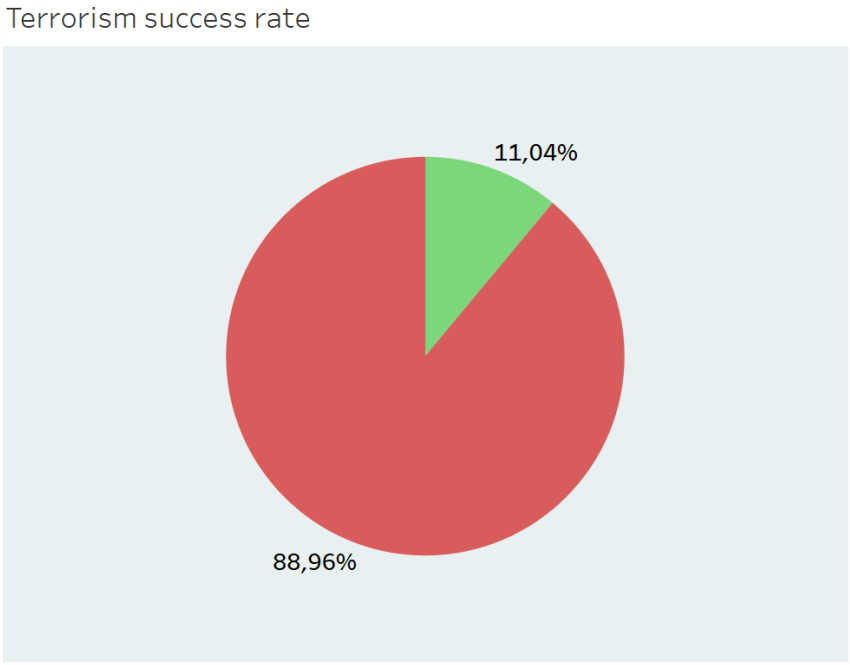


3.12 Regional analysis of cross-nationality attacks

The bubble chart below shows the distribution of the cross-nationality terror events by the regions. Cross-nationality terror events mean the terrorists and the targets have different nationalities. For this plot, the "Int Ideo" column is used. The plot indicates that most of the cross-nationality terror events happened in Eastern Europe, Australasia & Ocenia, and Middle East & North Africa regions.

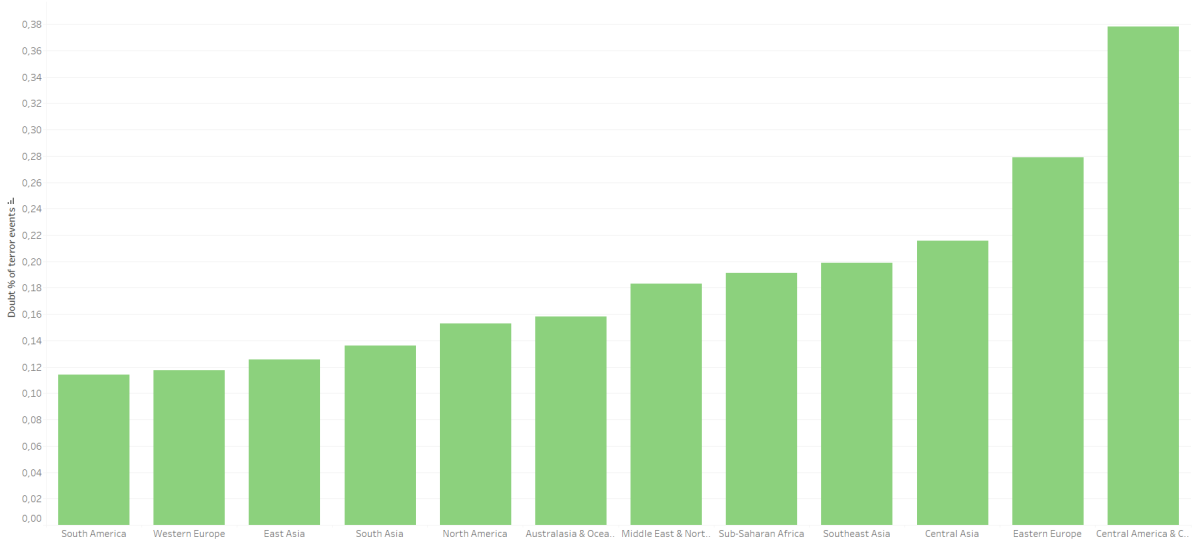
3.13 Success rate of all attacks

In the figure below, a pie chart is given. In this chart, red color displays attacks that are evaluated as successful and green color as failed. When cursor hovers over a color, tooltip is configured to show information for that color. Unfortunately it is seen that most of the attacks recorded (88,96%) happened successfully.



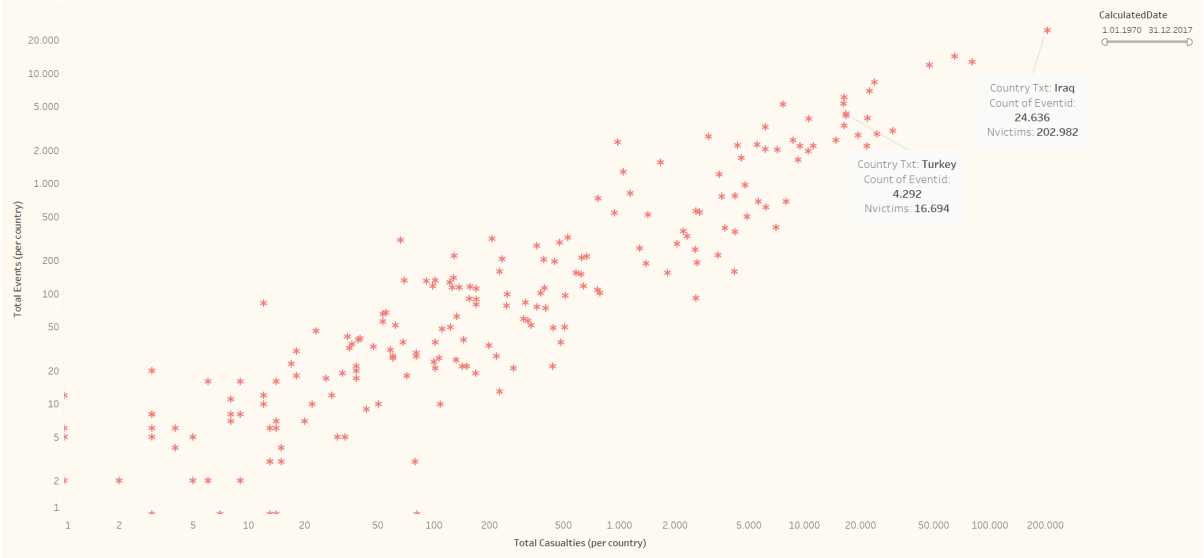
3.14 Bar plot analysis of doubt on terrorism element

The plot below shows the doubt about terror events in different regions. More than 28 percent of the attacks are doubtfully a terror event in Central America and Eastern Europe. Less than 12 percent of the events recorded in South America and Western Europe are doubtful. South American and Western European people are pretty sure that these attacks are caused by terror.



3.15 Country Based 'Suffer From Terror' Scatter Plot

The scatter plot below shows the total events on the y-axis and the total casualties on the x-axis. Each dot represents a country. The countries at the upper right side of the plot are the ones that suffer from terrorism the most. Iraq is the worst in both of these metrics and Turkey is also in a bad place sadly.



3.16 Heatmap analysis of categories

The heatmaps below show the relationship between attack types and target types. The one above examines this relationship by the number of events, and the below one examines this relationship by the number of deaths. We see that bombings and armed assaults are the most common attacks, and private citizens, military, and police are the people that are targeted most again. As different relationships we learn here, hostage attacks and hijackings mostly target private citizens.