Predictive Modelling of Terrorist Attacks Using Business Intelligence

Cornateanu Laurentiu ID10105187

QHO634

Supervisor Rahmat, Roushanak

Dissertation Project submitted in fulfilment of the requirement for the Degree of Bachelor of Science in Computer Science

September 28, 2022

Abstract

Terrorism employs aggression or the threat of violence and aims to instil fear, not only for the immediate victims but also among a broad audience.

Terrorism is distinguished from both conventional and guerrilla warfare by its reliance on fear. This project use technology to better understand what might also be applied to public protection, particularly in forecasting terrorist activity. Counterterrorism is a critical component of the worldwide authority framework, a long-term problem ensuring world protection innovation. Forecasting artificial intelligence in counterterrorism is presumed to be detrimental to human rights. This study is about transitioning from procedure to efficiency using Business Intelligence. It provides an integrated view of data information and historical, existing, and correlational insights that convert raw numbers into new initiatives. Business intelligence boosts your productivity by allowing you to address problems as they arise. Predictive analytics can quickly identify specific issues and enable targeted improvements.

“Do not fear those who operate in the shadows.”

— Michael Brady, The Fever

Table of contents

List of Figures and Tables

1. **INTRODUCTION**

I chose this data set for its importance of it. is obvious: the choice between death or life. It is not peace nothing beautiful in this world can happen: love, science, progress. Terrorism deterrence as a strategic goal of counterterrorism efforts what the purpose of this project. To predict the possible real-time risk of terror attacks around the Globe with the red flags in specific areas and make earlier high-impact triggers more visible using different models and measure the risk with Business intelligence. This research illustrates the ability of theoretically grounded models to predict and explain complex types of political violence at scales important to policymaking. Fewer to a nun terrorist attacks and victims of terrorism!

- This is the principals of the scope of this project.

1.1 **Background**

Artificial intelligence has recently gained traction, facilitating intelligent and automated decision-making across deployment scenarios and application areas. We are observing the convergence of several technologies (e.g., the Internet of Things, robotics, sensor technologies, etc.) and an increasing amount and variety of data and its new characteristics (e.g., distributed data) to use Artificial intelligence to scale. In data analytics and cyber security, Artificial intelligence can be seen as an emerging approach. Consequently, Artificial intelligence techniques have been used to support and automate relevant operations, e.g., traffic filtering and automated forensic analysis. Artificial intelligence and its application to, for instance, automated decision-making — particularly in safety-critical deployments such as autonomous vehicles, smart manufacturing, eHealth, etc. — can expose organisations and individuals to new and sometimes unpredictable risks and recent open attacks.

Artificial Intelligence provides businesses with unprecedented opportunities but also an incredible burden. Direct cause on people's lives has elevated significant concerns about Artificial intelligence ethics, data governance, confidence, and legitimacy. To establish trust, institutions must go far above identifying personally accountable BI fundamentals and put them into action. Accenture’s 2022 Tech Vision research found that.

1.2 **Problem Definition**

“Only 35% of global consumers trust how organisations implement Artificial Intelligence. and 77% think organisations must be held accountable for their misuse of Business intelligence.”

All publications and experts in the field talk about business intelligence transparency and fighting bias. Some others talk about the four and five pillars of Legally liable Business intelligence. Some about them showed little value had been put on hazard justification, including avoidance of reputational damage. It creates and encourages an executive philosophy that enables individuals to raise suspicions or concerns about artificial intelligence structures, creating individual or group pressure which is not correct. Now the question here is only one:

Does Artificial Intelligence take the facts and data from reality? And give only a reflection of that reality? Nothing more? The answer is YES! and more than this, BI have a tremendous power of truth and requires greater responsibility and strong leadership.

1.3 **Aim/Objectives of the project**

All continental regions identified by the Global Terrorism Database are included in our analysis of terrorist incidents committed between 1970 and 2020, classified into 13 geographic areas.

We can identify, evaluate, and compare the relevance of the main terrorism drivers across various geographical settings by specifically building predictive models, reducing the computing burden.

Additionally, by training machine learning models for distinct segments, the algorithms can choose various model parameters for multiple locations. All populated regions where terrorist activities may occur and be reported with high precision.

The research supports a classification that emphasises attacks against civilians, although it calls for changes when doing the analysis.

As a result, we first collect information on terrorism from the GTD dataset for each region from 1970 to 2020 and modify it to fit our classification into the project.

Then, utilising the different algorithms that provide a variety of geographic and socioeconomic characteristics worldwide, we partition each region.

Third, we extract from every unit in every region of the world:

The occurrence of terrorism using Artificial Intelligence data (1970–2020) and feature values from a collection of predictors collected at distinct geographical delays

The conceptual understanding of terrorism as a tactic informs the selection of the structural and procedural aspects included in the prediction model and has found several factors logically linked to the probability of its occurrence.

The project is based on addressing the questions about the threat posed by those who support terrorism and the ideological challenge it represents and seeing what success they have.

1.4 **Deliverables**

This research study proposes Predictive Modelling of Terrorist Attacks to provide a better sophisticated Artificial Intelligence mechanism. Most nations prioritise avoiding terrorist attacks above responding to them.
As a result,

Prediction is already essential to terrorism effectiveness.AI enables more significant amounts of information to be studied and may detect trends in those data that would be impossible to deauthorise because of their number and complexity. Complexity, without which interpretation by humans would be impossible. The consequence is that typical investigative technique that moves outward from identified individuals may be inadequately
complemented by approaches that assess the activities of a big part of a whole society to discover previously unidentified hazards. Even while the total number of far-right assaults remains modest compared to other types of terrorism, the rise in far-right political terrorism is one of the most alarming developments of the previous five years.

Since 2014, far-right assaults in North America and Western Europe have climbed by 250 per cent, while the death toll has increased by more than 700 per cent. In 2019, far-right terrorists were responsible for more than 80 murders, 50 of which occurred in the mosque assaults in New Zealand. Over the last five years, over 30 far-right terrorism events annually throughout the West.

Due to the enormity of the losses that might result from terrorist attacks, the Artificial Intelligence models created must be as accurate as possible to forestall many of them. A model's accuracy is measured by how well it can estimate the fraction of weak cells that have faced terrorism relative to the total number of positive instances anticipated.

It is still challenging to get a high degree of accuracy, particularly in areas with a low rate of terrorism. The proportion of positive instances relative to the total number of cases (both positive and negative) is the definition of prevalence.

2. Interrelationship between data and artificial intelligence

2.1 Global Terrorism Database

This research is using “https://www.start.umd.edu/gtd/”

“The Global Terrorism Database™ (GTD) is an open-source database including information on terrorist events around the world from 1970 through 2020 (with annual updates planned). Unlike many other event databases, the GTD includes systematic data on domestic and international terrorist incidents that have occurred during this period and now includes more than 200,000 cases.”

For this research and analyse, the use of 'globalterrorismdb_0522dist.csv.'

2.2 Artificial Intelligence (AI)

As a subcategory of machine learning, this technology for artificial intelligence alters our conception of the correlation between decision-making and analytics. Rather than teaching the machine how to understand, deep knowledge allows the data to teach the computer, resulting in more accurate prediction models with each new batch of data provided to the computer.

2.2.1 Constraints

The limitation is the current understanding of AI. The perspective is the same as that of software engineering. At its essence, AI is mathematical. It is only a collection of procedures that accept input parameters and produce output variables. The next step is to interpret the results. You must ensure that the approach you choose applies to the region in which you will conduct research and make forecasts. Ensure the assumptions of this technique correspond well with the problem space's beliefs. You cannot arbitrarily apply any random approach to any specific issue, and the "mystical AI button" will not make it happen. Unfortunately, the system does not operate in this manner.

Artificial intelligence is an artistic expression since it involves mathematical concepts and depends on how a person employs mathematical tools. This art form is still evolving. We are attempting to design the future without a well past. Currently, we are still learning as we go. There have been a few examples of success over the years.

2.2.2 Benefits

Artificial intelligence and machine learning provide the basis of a new managerial strategy that now has the capabilities to engage this digital frontier, from combating cyber dangers to increasing consumer advertising. According to research, fifty-five per cent of businesses have used AI in at least one company function. Machine-learning and artificial intelligence has the potential to add hours to your routine if you understand how to use them effectively.

As a subcategory of machine learning, this technology for artificial intelligence alters our conception of the correlation between decision-making and analytics. Rather than teaching the machine how to understand, deep learning allows the data to teach the computer, resulting in more accurate prediction models with each new batch of data provided to the computer.

2.3 Why Predictive Modelling of Terrorist Attacks Using Business Intelligence?

Prescriptive modelling is a collection of methods used to draw predictions about future occurrences that are unclear. In the academic arena, one may be interested in predicting a measurement of learning teaching or other proxy metrics of value for organisations that already include predictive analytics. In interpretative analysis, the objective is to explain a particular result using all available facts.

Terrorism is intentionally unexpected. The provocation of fear relies on the idea that an assault will be carried out by an unknown individual at a random location and time. Nevertheless, successful counterterrorism strategies are supported by the potential for prediction. The issue is not if prediction has a role in fighting terrorist acts, to whether and in what ways specific AI-based strategies might improve prediction.

2.4 Characteristics of the data set

The dataset used for this project has included statistics on more than 91,000 explosions, 20,000 assassinations, and 13,000 terrorism-related deaths.

• The database comprises 135 columns and 181,692 records detailing terrorist acts

These are the data entries we are working with:

- ✓ Event ID
- ✓ Year of aggression
- ✓ Numerical code for the Country
- ✓ Name of the Country
- ✓ Method of assault
- ✓ Type of objective
- ✓ Subtype of the target
- ✓ Identifying target
- ✓ Gang name
- ✓ Subtype of firearm
- ✓ Quantity of deaths (Numeric value of the number of people killed in the attack)
- ✓ The worth of the property (Numeric value of the loss of property due to the attack)
- ✓ Criticality of attack (Criticality is "1" if the attack causes significant loss of life and property, and "0" if the assault is not overly critical)
- ✓ Percentage of successful attacks

3. Methodology

The motivating challenge for us has been to comprehend the dataset, which has led us to analyse how terrorist activity has impacted people's lives and, doing that to a better understanding, divide it into two parts questions: Visualisation and Predictions. Where each one has a different strategy to analyse the data set that is materialised in two assorted designs by the questions asked and the perspective.

3.1 The approach

The main aim of this chapter is to provide an approach to the solution or to consider specific criteria for the solution. This chapter explains the approach taken in the evaluation process and techniques used to carry out the evaluation and the results of the evaluation

3.1.1 Visualisation

Quantitative Analysis of Global Terrorist Attacks Based on Machine Learning

In this project, for visualisation, we will be answering 12 questions

- ✓ What year data consist of terrorist activities ranging from the year to what year
- ✓ Maximum number of people killed in an event
- ✓ Maximum number of people wounded in an event
- ✓ Max number of total casualties in an event
- ✓ Relationships between quantitative variables or categorical variables
- ✓ Terrorist Activities by Region in each Year
- ✓ Several off attacks were there in 1970 & 2020, and the percentage of attacks has also increased.
- ✓ Total casualties by wounded + killed each year
- ✓ Attack by Country
- ✓ Method of attack
- ✓ Type of target
- ✓ The most active terrorist organisation

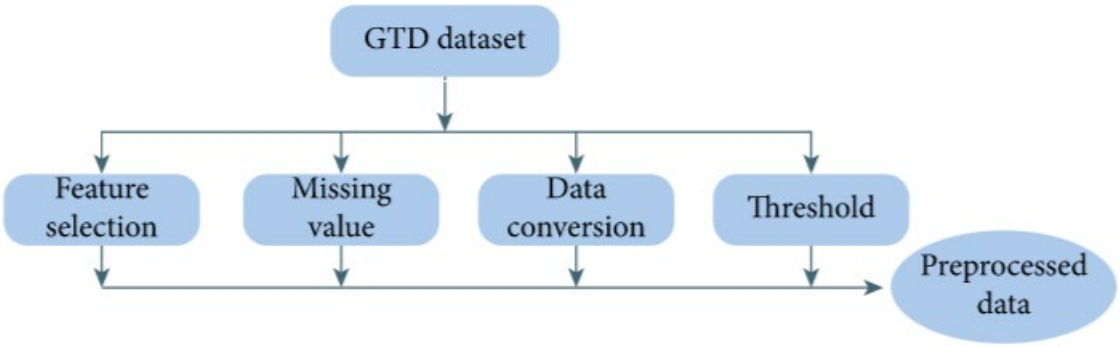

Figure 1. Cycle Processing Visualising

3.1.2 Predictions

In this part of the project, we will be answering these 5 points and involved in a prediction:

- ✓ Most terrorism affects countries each decade.
- ✓ To find the most popular target type assassinated
- ✓ The most active terrorist group activity over the years was weapons used and kills caused.
- ✓ Economical loss of each decade and its comparison based on the criticality of the attack.
- ✓ Regression to see the success of the terrorist attack
- ✓ To predict the success of terrorist activities based on multiple factors

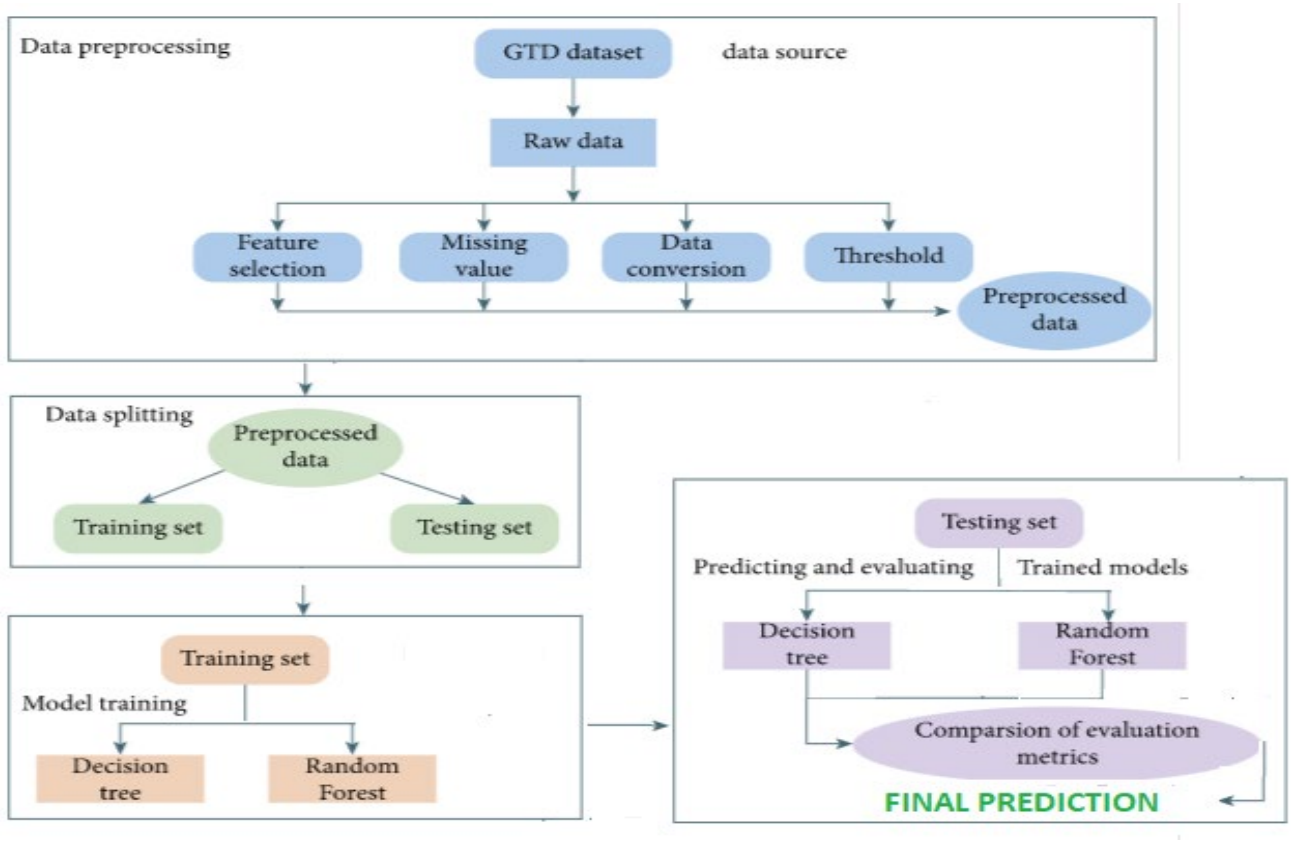

Figure 2. Cycle processing Prediction

3.2 Evaluation criteria

Concentrating on providing an estimate of the number of terrorist events that will occur in the following years so that the authority in charge may take the necessary steps to limit the number of occurrences—also, picturing prior terrorist incidents to comprehend the pattern, as well as to determine which nations are prone to assaults. I am eager to address the difficulty of identifying the terrorist organisation behind the assault.

3.2.1 Constraints

• The data contains mappings from non-numeric to numeric entries for most characteristics.
• Despite this, it does not provide such a mapping for "gname". A new feature called "gno"is introduced for mathematical modelling reasons.
• Between 1980 and 1990, the dataset was not kept for a brief period.
• The unknown values have been filled in differently in several locations. "unknown" in "property" is populated with "-9", but "unknown" in "nperps" is populated with "-99."
• A set of features introduced after 1997 lack information for the years before 1970, and there are a few features withdrawn after a particular year for which there is no information since then.
Attempts have been made to address these gaps via further collecting activities.
Consequently, the dataset has several missing values.

3.3 The approach Software & application use

The software used for the project is Anaconda

"Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment." https://www.google.com/search?client=firefox-b-d&q=why+use+anaconda

3.3.1 Why Anaconda!?

Provides all the common tools used in computer science in a single package, eliminating the need to install each item separately with its dependencies. Aside from the convenience, there will be no significant difference between utilising a gnosis anaconda and creating your own ecosystem. Under the engine, they are all the same python.

3.3.2 Why Jupiter notebook applications?

Jupyter notebooks offer an interactive computational environment that enables the creation of applications in the field of data science. Jupyter notebooks integrate software code, computational output, explanatory prose, and substantial content into a single document. Notebooks provide the modification and execution of code in a web browser and show the results of computations.

3.4 Summary

This chapter explains the approach taken in the evaluation process and techniques used to carry out the evaluation and the results of the evaluation. The importance of the evaluation phase lies in identifying the level of importance of the problem, the uniqueness, and effectiveness of the approach adopted to solve the problem, and the usefulness, usability and quality of the solution provided.

And discusses the evaluation of the research carried out. It also details the approach adopted in the Project to evaluate the proposed solution. Critically evaluate the solution.
Many vital facts regarding the features and drawbacks of the system were discovered during the evaluation process.

4. Analysis of data

This chapter aims to provide an overview of the systems development methodology selection. This discusses in detail the selection of the development methodology and the modelling and investigates the available development methodologies and evaluates them before suggesting the suitable methodology for the proposed solution. Furthermore, this chapter will detail the technical aspects of the research. Explore various models and approaches related to the proposed solution, then critically evaluate them, and decide on the right approach to be adopted.

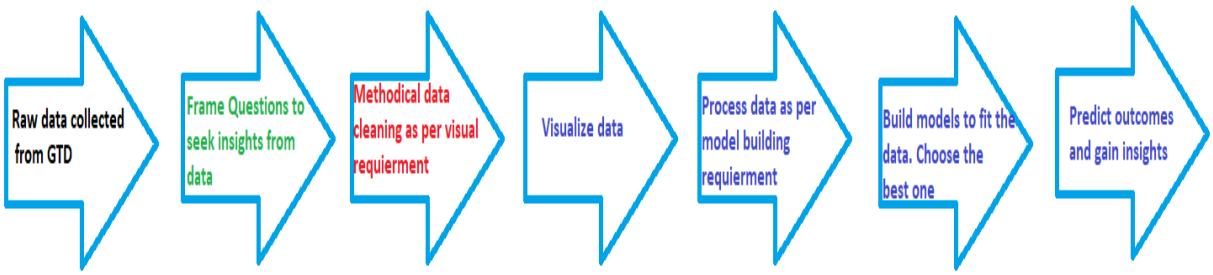

Figure 3. Methodology Progress

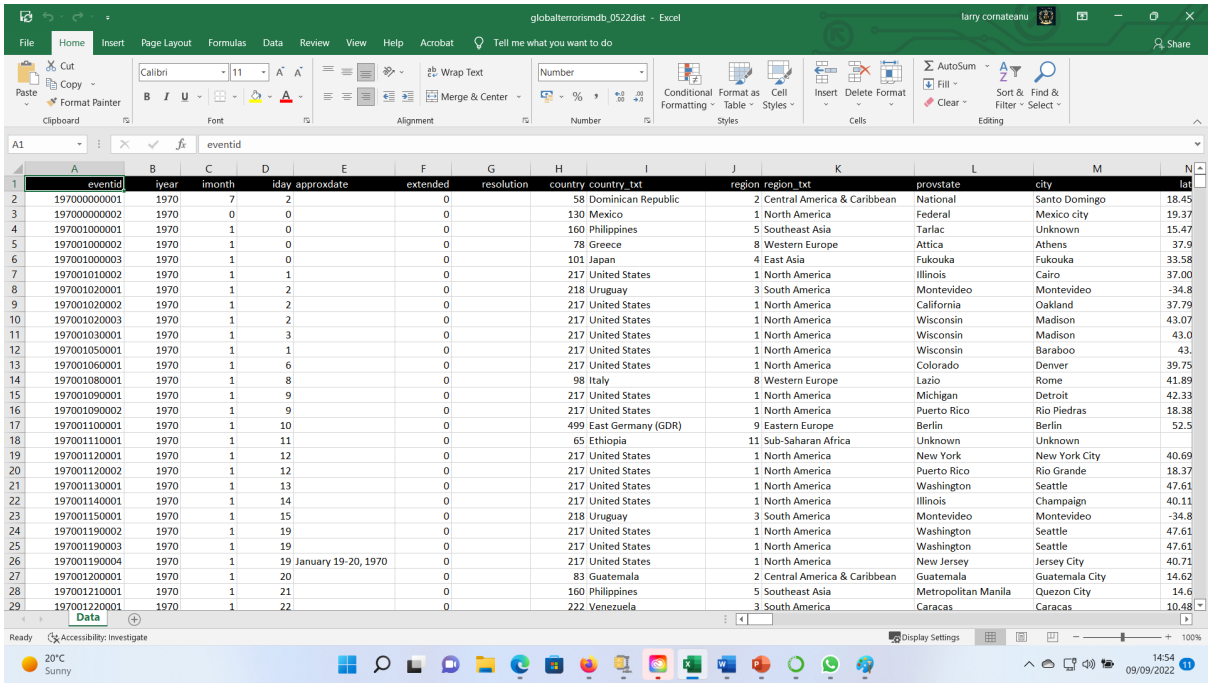

Figure 4. GDT data "globalterrorismdb_0522dist.csv"

4.3 Data Cleaning

The dataset has 145 characteristics with several missing values. Some of the features have been added over time. The difficulty is comprehending each property's significance and

providing meaningful values to every attribute's incomplete data. Various kinds of visualisations need distinct data purification procedures. Because undetermined values, not accessible values, and preliminary data have abnormal meanings in many attribute circumstances, particularly from the standpoint of global terrorism, we cannot simply eliminate them. We meticulously imputed missing data using replacements such as means, zeros, constants, etc.

To demonstrate all three types of imputations, the following examples are provided. To replace incomplete data in the Number of kills(nkill) property, we calculated the means of the Number of kills per weapon. We substituted those values for the relevant missing values. To fill blank values in the Number of perpetrators (nperp) column, I used 9. When no ransom was sought, missing ransom amount variables were filled with 0. Specific properties had to be translated to numeric form to perform mathematical operations.

As there were several qualities, the pre-processing of information required Principal Component Analysis to identify the most critical Principal Components.

Consequently, the dataset has several missing values that must be addressed.

Our database includes 135 categories we will edit in compliance with the statement's requirements. We will sanitise our data depending on each related to the point in time needs, retaining just the needed info and removing the rest.

4.4 Summary of the review of the literature

The review includes around four research articles published in IEEE publications.

This was the only one of the four publications that resembled my approach to tackling the issue. In this project, a Terrorist group prediction model was developed, and it analyses the trend of the kind of attacks carried out by various terrorist groups over time. Target victim analysis, the level of damage to public property, the type of the assault, the weapons used, and the location of the attack. The terrorist group prediction model can learn patterns to anticipate which group may have begun a particular assault. The partition-based clustering method was applied to create a Terrorist group prediction model. The technique was used as it is effective with categorical characteristics and missing data.

Using data mining prediction methods such as the Terrorist group prediction model and cluster-based methodologies, this research concludes that historical data may be utilised to forecast the terrorist organisation responsible for a specific incident using past data.

5. Unsupervised machine learning

The cornerstones of early data analysis are visualisations. Initially, to gain a feel of the dataset, we produced simple visualisations. As analysts, this enables us to comprehend what the data truly represents, i.e., the relevance of the data.

Through data visualisation, patterns, trends, and correlations that could go unnoticed in text-based information can be highlighted and identified more easily.

5.1 Visualisation

The cornerstones of early data analysis are visualisations. To get an initial understanding of the data, I created simple visuals. As an analyst, this enables us to comprehend the relevance of the data, to comprehend the meaning of the data. Through data visualisation, patterns, trends, and correlations that could go unnoticed in text-based data can be highlighted and identified more easily. Subletting with distinct characteristics allows us to examine the database based only on that characteristic.

Example:

Year-by-year analysis, month-by-month analysis, target-by-target analysis, attack type-by-attack type analysis, and primary weapon type-by-type analysis.

These packages have been imported to better analyse the data:

5.1.1 Import packages in Python

Integration is the process of granting a Python file or package permission to a module inside such a Python file or unit. Only functions and attributes that your application can access may be used. To access mathematical functionality, for instance, you must load the math package first.

data processing

- import pandas as pd
- linear algebra
- import numpy as np

data visualisation

- import matplotlib.pyplot as plt
- matplotlib inline
- import seaborn as sns
- import plotly. express as px
- import dash
- import dash_core_components as dcc
- import dash_html_components as html
- import squarify
- from matplotlib import pyplot as plot

5.1.2 import dataset from GTD

The information includes the number of columns, column names, column data types, memory use, range index, and the number of cells in each column (non-null values).
” https://www.python.org/”

Data has imported

	eventid	iyear	imonth	iday	approxdate	extended	resolution	country	country_txt	region	...	addnotes	scite1	scite2	scite3	dbsource	INT_LOG
0	197000000001	1970	7	2	NaN	0	NaN	58	Dominican Republic	2	...	NaN	NaN	NaN	NaN	PGIS	0
1	197000000002	1970	0	0	NaN	0	NaN	130	Mexico	1	...	NaN	NaN	NaN	NaN	PGIS	0
2	197001000001	1970	1	0	NaN	0	NaN	160	Philippines	5	...	NaN	NaN	NaN	NaN	PGIS	-9
3	197001000002	1970	1	0	NaN	0	NaN	78	Greece	8	...	NaN	NaN	NaN	NaN	PGIS	-9
4	197001000003	1970	1	0	NaN	0	NaN	101	Japan	4	...	NaN	NaN	NaN	NaN	PGIS	-9

5 rows × 135 columns

Figure 5. Importing data

5.1.3 Dataset information

The data includes the number of columns, column labels, column data types, memory consumption, range index, and the number of cells in each column (non-null values). The info () method really outputs the information.” https://www.python.org/”

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209706 entries, 0 to 209705
Columns: 135 entries, eventid to related
dtypes: float64(54), int64(23), object(58)
memory usage: 216.0+ MB
```

Figure 6.Data Information

5.1.4 The shape of a DataFrame is a tuple of array dimensions that tells the number of rows and columns of a given DataFrame.” https://www.python.org/”

```
df.shape
```

```
(209706, 135)
```

Figure 7. Data Frame number of rows and column

5.1.5 The DataFrame form corresponds to the number of rows and columns.

```
df.columns
```

```
Index(['eventid', 'iyear', 'imonth', 'iday', 'approxdate', 'extended',
       'resolution', 'country', 'country_txt', 'region',
       ...
       'addnotes', 'scite1', 'scite2', 'scite3', 'dbsource', 'INT_LOG',
       'INT_IDEO', 'INT_MISC', 'INT_ANY', 'related'],
      dtype='object', length=135)
```

Figure 8. Columns labels

5.2 Cleaning the data

Data analysts devote a substantial amount of effort to cleaning datasets and preparing them for use. Data Scientists must have the ability to cope with untidy data, missing values, inconsistent, noisy, or incomprehensible data. Python has a built-in

	iyear	imonth	iday	country_txt	region_txt	provstate	city	latitude	longitude	location	summary	attacktype1_txt	targtype1_txt	gname	w
0	1970	7	2	Dominican Republic	Central America & Caribbean	National	Santo Domingo	18.456792	-69.951164	NaN	NaN	Assassination	Private Citizens & Property	MANO-D	
1	1970	0	0	Mexico	North America	Federal	Mexico city	19.371887	-99.086624	NaN	NaN	Hostage Taking (Kidnapping)	Government (Diplomatic)	23rd of September Communist League	
2	1970	1	0	Philippines	Southeast Asia	Tarlac	Unknown	15.478598	120.599741	NaN	NaN	Assassination	Journalists & Media	Unknown	
3	1970	1	0	Greece	Western Europe	Attica	Athens	37.997490	23.762728	NaN	NaN	Bombing/Explosion	Government (Diplomatic)	Unknown	
4	1970	1	0	Japan	East Asia	Fukouka	Fukouka	33.580412	130.396361	NaN	NaN	Facility/Infrastructure	Government	Unknown	

Figure 9. Removal of errors

5.3 Rename the column

Occasionally it is straightforward and appropriate and more understandable to rename the column. By convention, the renaming function returns a new Python data structure with modified column and row labels. As stated before, this implies that renaming will keep the original data frame unaltered by default. If in place = True is used, rename will not generate new output." https://www.python.org/"

	Year	Month	Day	Country	state	Region	city	latitude	longitude	AttackType	Killed	Wounded	Target	Summary	Group
0	1970	7	2	Dominican Republic	National	Central America & Caribbean	Santo Domingo	18.456792	-69.951164	Assassination	1.0	0.0	Julio Guzman	NaN	MANO-D
1	1970	0	0	Mexico	Federal	North America	Mexico city	19.371887	-99.086624	Hostage Taking (Kidnapping)	0.0	0.0	Nadine Chaval, daughter	NaN	23rd of September Communist League
2	1970	1	0	Philippines	Tarlac	Southeast Asia	Unknown	15.478598	120.599741	Assassination	1.0	0.0	Employee	NaN	Unknown
3	1970	1	0	Greece	Attica	Western Europe	Athens	37.997490	23.762728	Bombing/Explosion	NaN	NaN	U.S. Embassy	NaN	Unknown
4	1970	1	0	Japan	Fukouka	East Asia	Fukouka	33.580412	130.396361	Facility/Infrastructure Attack	NaN	NaN	U.S. Consulate	NaN	Unknown

Figure 10. Rename the Column

5.4 Create a new column

Create a new column to assist us in better comprehending our research project analysis "Killed"+" Wounded"="Casualty"

" https://www.python.org/"

	Year	Month	Day	latitude	longitude	Killed	Wounded	Casualty
count	209706.000000	209706.000000	209706.000000	205015.000000	205014.000000	209706.000000	209706.00000	209706.000000
mean	2004.800993	6.455285	15.527930	23.358696	30.416738	2.285810	2.79251	5.078319
std	13.519321	3.387098	8.801104	18.137061	56.113029	11.012018	38.93325	44.832867
min	1970.000000	0.000000	0.000000	-53.154613	-176.176447	0.000000	0.00000	0.000000
25%	1992.000000	4.000000	8.000000	11.510046	8.748117	0.000000	0.00000	0.000000
50%	2012.000000	6.000000	15.000000	31.300213	43.746215	0.000000	0.00000	1.000000
75%	2015.000000	9.000000	23.000000	34.557022	68.835918	2.000000	2.00000	4.000000
max	2020.000000	12.000000	31.000000	74.633553	179.366667	1700.000000	10878.00000	12263.000000

Figure 11. Create a new Column

5.5 Missing values + Dataset Information

Comes back to the number of variables that are missing from the data collection. Skipping rows with missing values is a straightforward method for dealing with data having missing values. Fig. 11 + The info() function outputs data about the DataFrame. The data includes the number of sections, section titles, section data formats, memory consumption, range index, and the number of cells within every section (non-null values).Fig.12

" https://www.python.org/"

```
data.isnull().sum()
Year              0
Month             0
Day               0
Country           0
state             0
Region            0
city            426
latitude       4691
longitude      4692
AttackType        0
Killed            0
Wounded           0
Target          635
Summary       66120
Group             0
Target_type       0
Weapon_type       0
Motive       154648
Casualty          0
dtype: int64
```

Figure 13. Number of NULL and NAN values

```
data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209706 entries, 0 to 209705
Data columns (total 19 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Year         209706 non-null  int64
 1   Month        209706 non-null  int64
 2   Day          209706 non-null  int64
 3   Country      209706 non-null  object
 4   state        209706 non-null  object
 5   Region       209706 non-null  object
 6   city         209280 non-null  object
 7   latitude     205015 non-null  float64
 8   longitude    205014 non-null  float64
 9   AttackType   209706 non-null  object
 10  Killed       209706 non-null  float64
 11  Wounded      209706 non-null  float64
 12  Target       209071 non-null  object
 13  Summary      143586 non-null  object
 14  Group        209706 non-null  object
 15  Target type  209706 non-null  object
 16  Weapon type  209706 non-null  object
 17  Motive       55058 non-null   object
 18  Casualty     209706 non-null  float64
dtypes: float64(5), int64(3), object(11)
memory usage: 30.4+ MB
```

Figure 12. Information about data

5.6 Statistical information

The method DataFrame.hist () is important for comprehending the distribution of numeric variables. This function divides the values into their respective numerical variables. Its primary purpose is to generate the Histogram of a given Data frame.

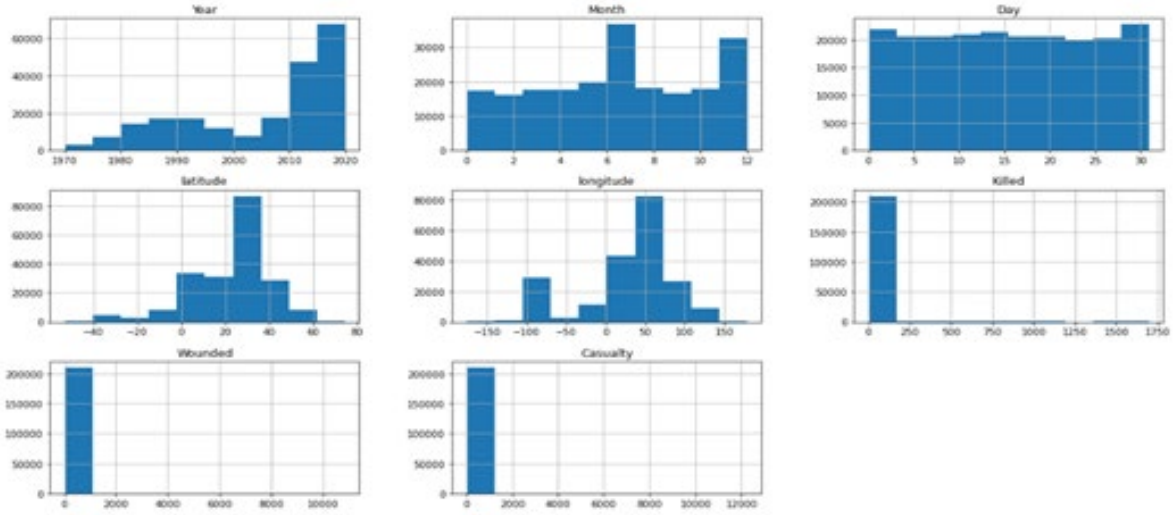

Figure 14. Distribution variables

5.7 Data Information

This function outputs information about a DataFrame, including the dtype and columns of the index, non-null values, and memory utilisation. DataFrame about which to output information. Whether to print the full summary." https://www.python.org/"

```
data.info
```

```
<bound method DataFrame.info of         Year  Month  Day
0       1970      7    2  Dominican Republic      National
1       1970      0    0              Mexico       Federal
2       1970      1    0         Philippines        Tarlac
3       1970      1    0              Greece        Attica
4       1970      1    0               Japan       Fukouka
...      ...    ...  ...                 ...           ...
209701  2020     12   31               Yemen    Al Hudaydah
209702  2020     12   31               Yemen    Al Hudaydah
209703  2020     12   31             Germany   Lower Saxony
209704  2020     12   31         Afghanistan         Kabul
209705  2020     12   31        Burkina Faso         Sahel

                          Region           city   latitude
0       Central America & Caribbean  Santo Domingo  18.456792
1                     North America     Mexico city  19.371887
2                   Southeast Asia         Unknown  15.478598
3                   Western Europe          Athens  37.997490
4                        East Asia         Fukouka  33.580412
...                            ...             ...        ...
209701   Middle East & North Africa           Sabaa  15.305307
209702   Middle East & North Africa    Beit Maghari  13.931337
209703                Western Europe         Leipzig  51.342239
209704                    South Asia           Kabul  34.523842
209705           Sub-Saharan Africa           Kelbo  13.864252

                           AttackType  Killed  Wounded  \
0                       Assassination     1.0      0.0
1          Hostage Taking (Kidnapping)     0.0      0.0
2                       Assassination     1.0      0.0
3                   Bombing/Explosion     0.0      0.0
4       Facility/Infrastructure Attack     0.0      0.0
...                                ...     ...      ...
209701                  Bombing/Explosion     0.0      0.0
209702                  Bombing/Explosion     0.0      0.0
209703  Facility/Infrastructure Attack     0.0      0.0
209704                  Armed Assault     1.0      0.0
209705                  Armed Assault     5.0      0.0
```

Figure 15. Information about Dataset

5.8 Describe the Data

Pandas describe () is used to display certain fundamental statistical features, such as the percentile, mean, and standard deviation, of a data frame or a sequence of numeric numbers. When applied to a string sequence, this technique provides a distinct result.

” https://www.python.org/”

```
data.describe()
```

	Year	Month	Day	latitude	longitude	Killed	Wounded	Casualty
count	209706.000000	209706.000000	209706.000000	205015.000000	205014.000000	209706.000000	209706.00000	209706.000000
mean	2004.800993	6.455285	15.527930	23.358696	30.416738	2.285810	2.79251	5.078319
std	13.519321	3.387098	8.801104	18.137061	56.113029	11.012018	38.93325	44.832867
min	1970.000000	0.000000	0.000000	-53.154613	-176.176447	0.000000	0.00000	0.000000
25%	1992.000000	4.000000	8.000000	11.510046	8.748117	0.000000	0.00000	0.000000
50%	2012.000000	6.000000	15.000000	31.300213	43.746215	0.000000	0.00000	1.000000
75%	2015.000000	9.000000	23.000000	34.557022	68.835918	2.000000	2.00000	4.000000
max	2020.000000	12.000000	31.000000	74.633553	179.366667	1700.000000	10878.00000	12263.000000

Figure 16. Statistical data description

5.9 Correlation Analysis

A statistical approach that demonstrates the relationship between variables. The dataframe.corr () method of Pandas is used to generate the correlation matrix. It is used to determine the pairwise correlation between each column in the data frame

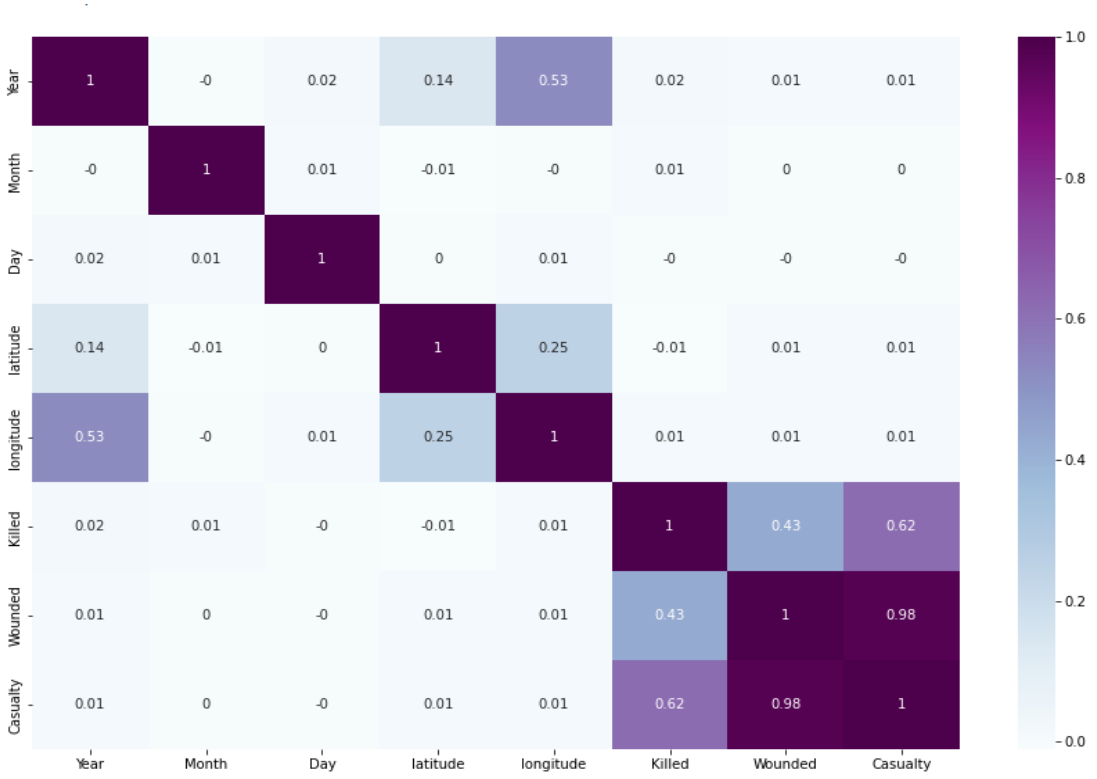

Figure 17. Correlation matrix

5.10 Graphic representation of terrorist activities by regions

Computes a basic cross-tabulation of two or more elements.

“Unless an array of values and an aggregation function are supplied, computes a frequency table of the factors by default.”

” https://www.python.org/”

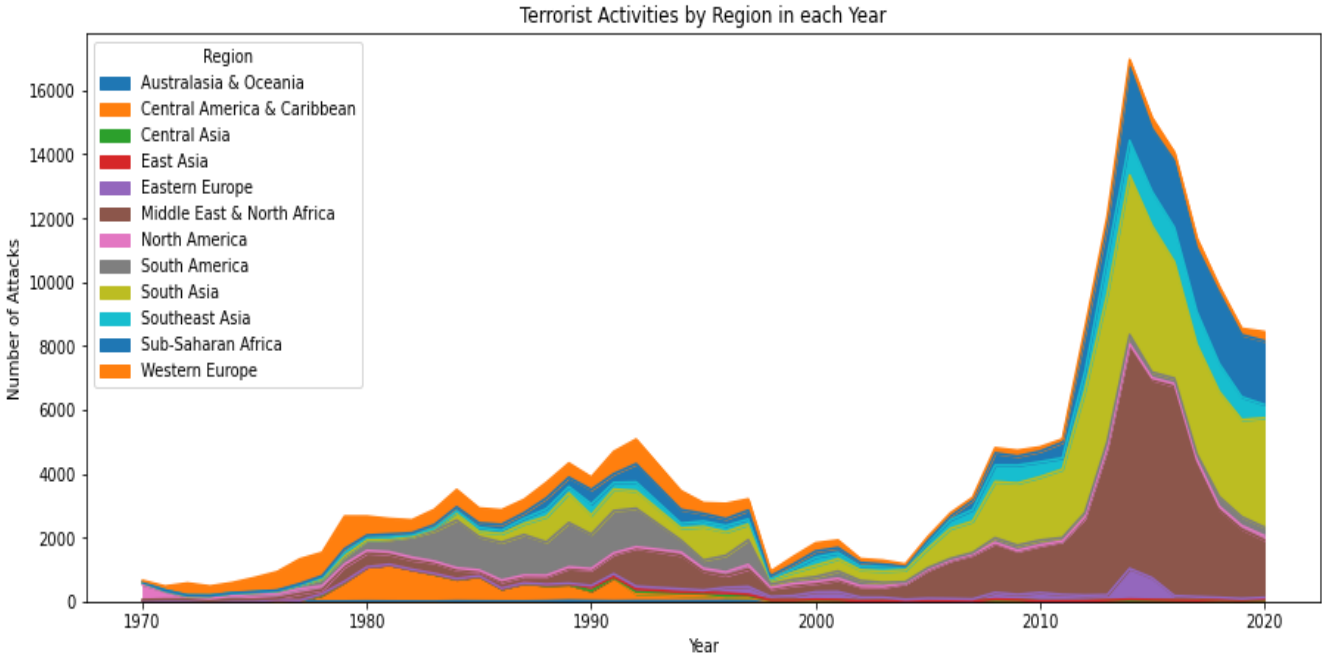

Figure 18. Numbers of terrorist Activities each year

5.11 Numbers of attack increases over time

The value counts () method provides objects containing counts of unique values.
The resultant object will be in decreasing frequency order, with the most common element appearing first.
https://www.python.org/

```
651 attacks happened in 1970 & 8438 attacks happened in 2020
So the number of attacks from 1970 has increased by 92.0 % till 2020
```

Number of attack were there in 1970 & 2020 and Also find percentage the attacks have increased.

Figure 19. Percentage increase of attack from 1970 to 2020

5.12 Total casualties by wounded + killed by year by Countries

"Dataset.groupby(). sum() is used to aggregate rows based on one or more columns, followed by the sum agg function. The groupby() method returns a DataFrameGroupBy object that includes the aggregate function sum() for calculating the sum of a specific column for each group. "

" https://www.python.org/"

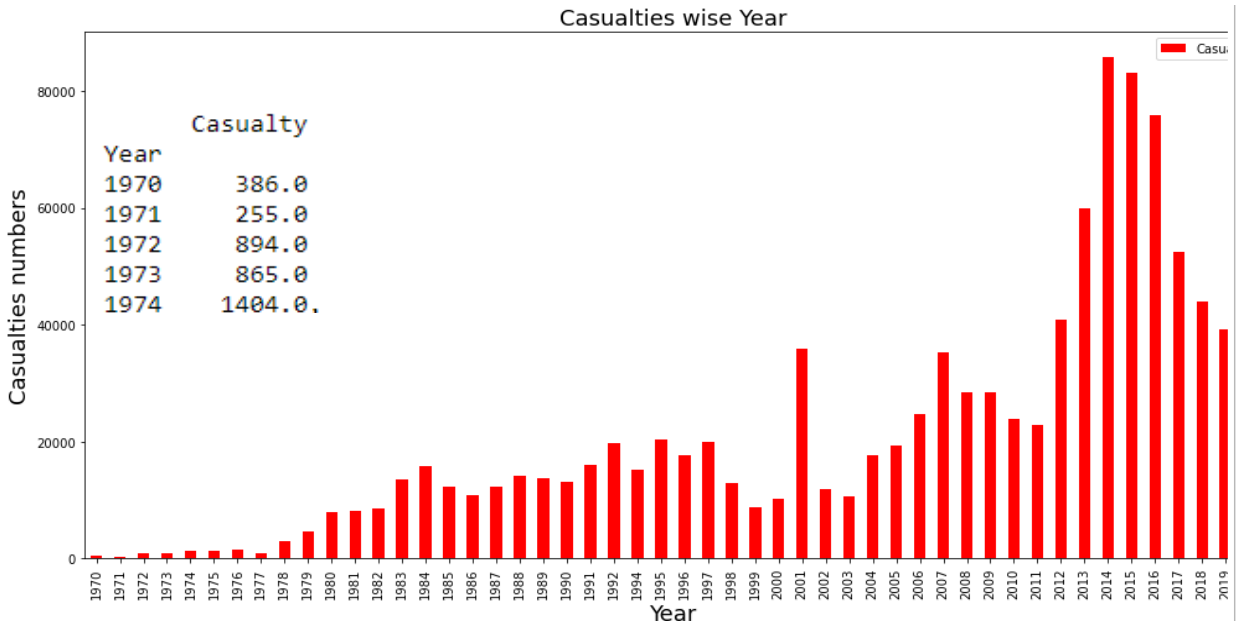

Figure 20. Casualties by year

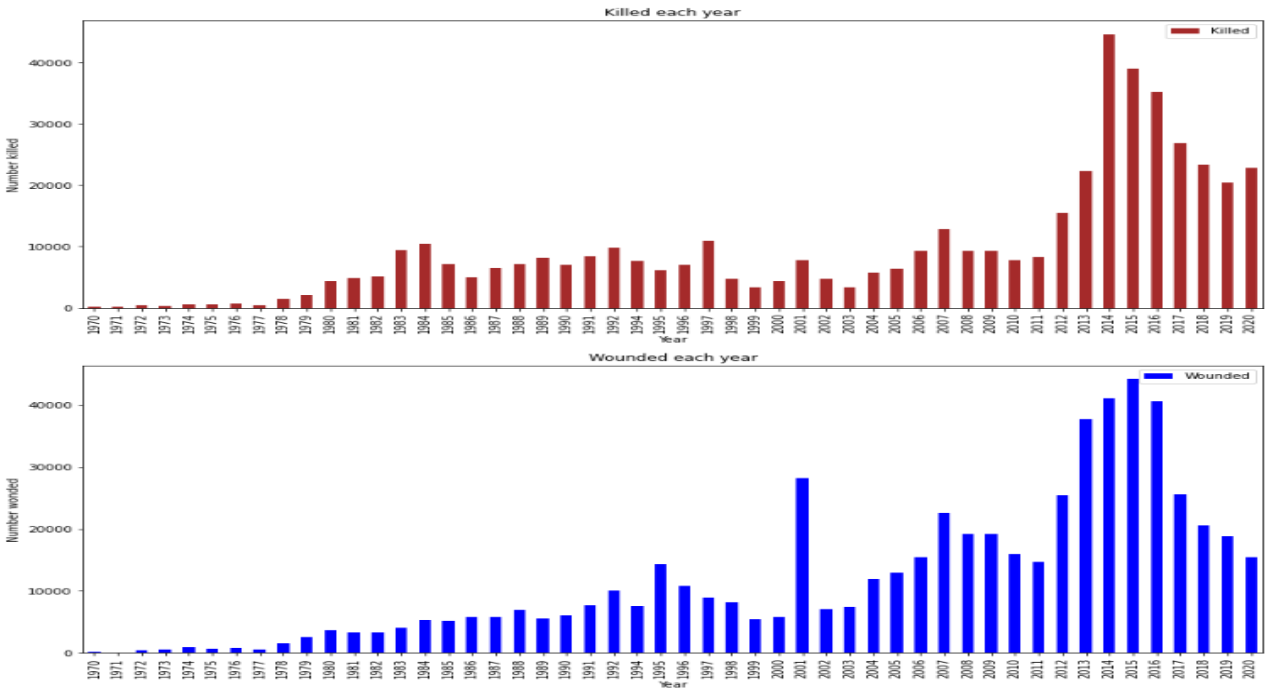

Figure 21. Killed & wounded each year

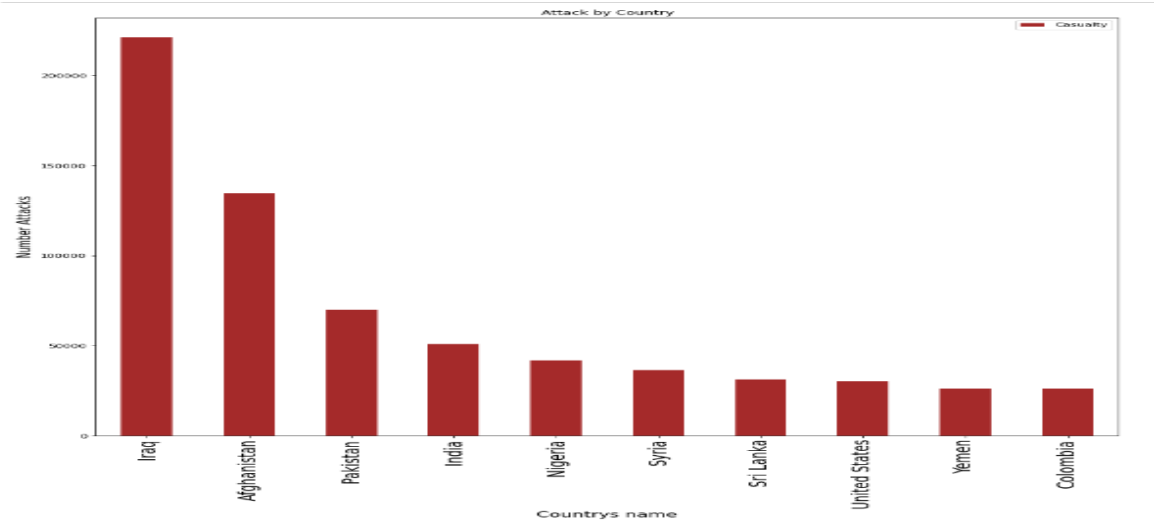

Figure 22. Attack by Country

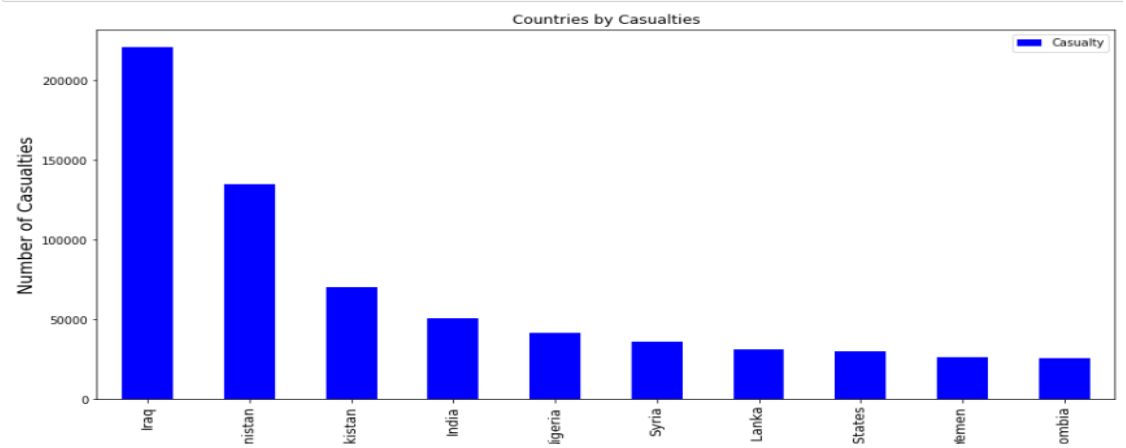

Figure 23. Casualties by Countries

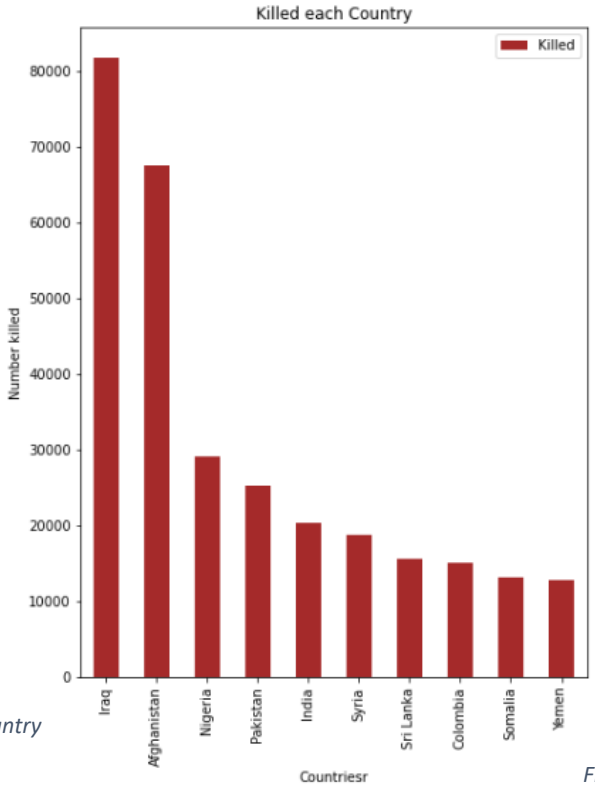

Figure 24. Killed by Country

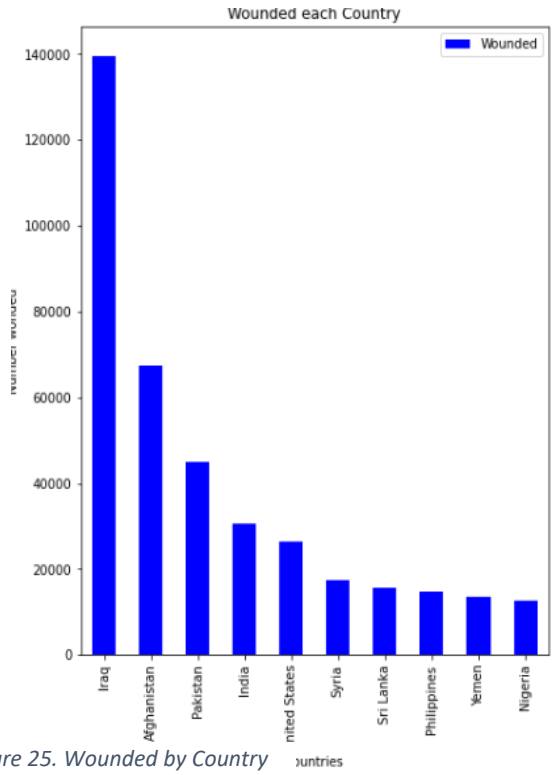

Figure 25. Wounded by Country

5.18 Number of Casualties vs killed for each Country by year (Animated)

DataFrame.plot.scatter

Each point's parameters are given by two data frame columns, and each point is represented by a filled circle. This kind of visualisation is effective for identifying intricate relationships between two variables. Points are natural 2D coordinates such as longitude and latitude on a map, or any pair of metrics that may be displayed against one another. This scatter plot can be visualised in the attachment sent with the project WinZip.

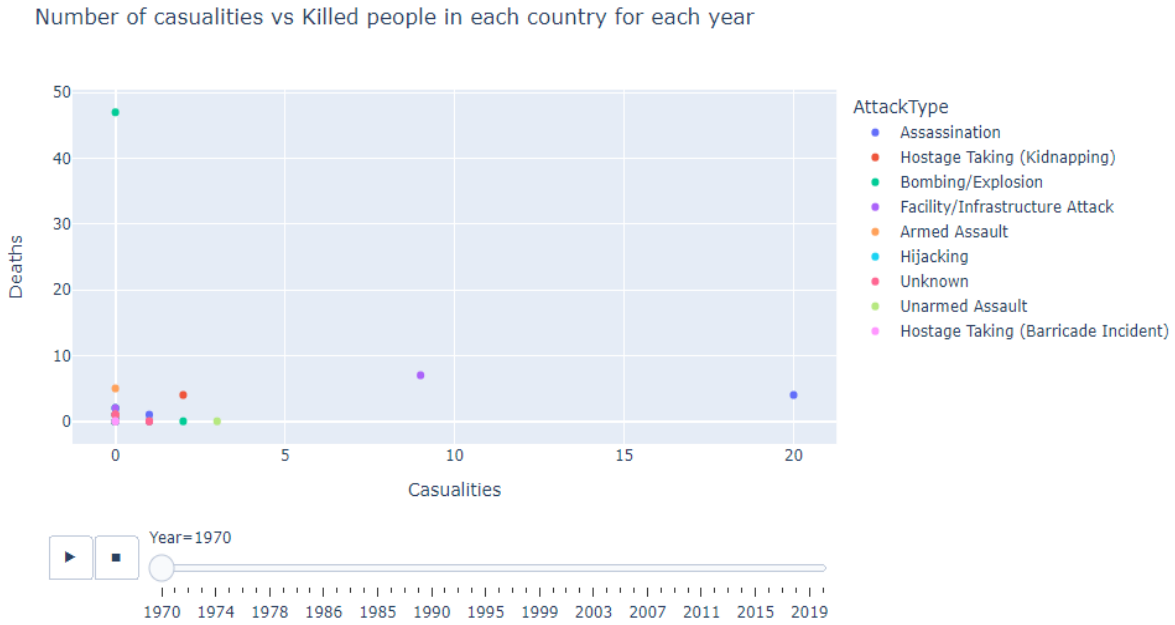

Figure 26. Animated Jupiter Notebook Number of Casualties vs Killed and Attack Type in each Country over the years

5.19 Number of Attacks in every Country by year (Animated)

A Choropleth Map is a map composed of coloured polygons. It is used to represent spatial variations of a quantity total number of attacks on each country. This Choropleth Map can be visualised in the attachment sent with the project WinZip.

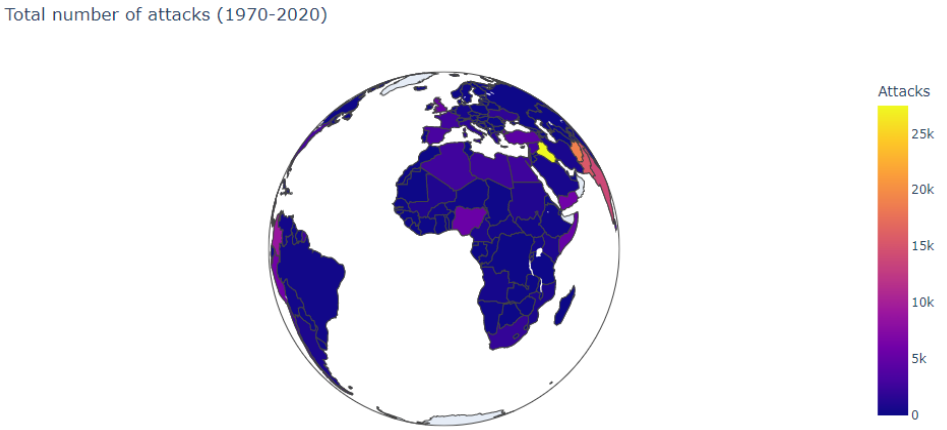

Figure 27. Total numbers of attacks in every Country between 1970-2020 (Animated)

5.20 Graphic Representation of numbers of casualties every year

Using Squarify to Plot a Treemap

As with any other chart type and data visualisation technique, treemap charts are only useful when their use is appropriate and justifiable, such as when comparing years.

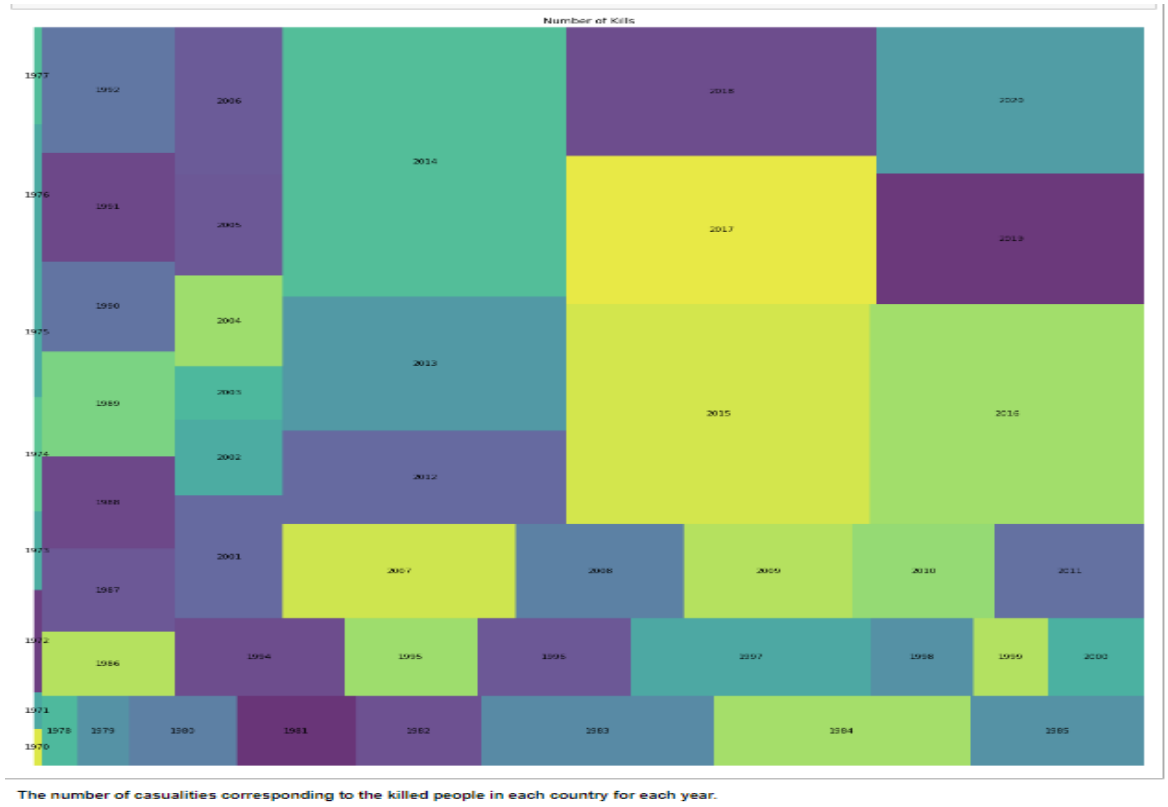

Figure 28. Number of casualties by each year 1970 – 2020

5.21 Analyse terrorist activities

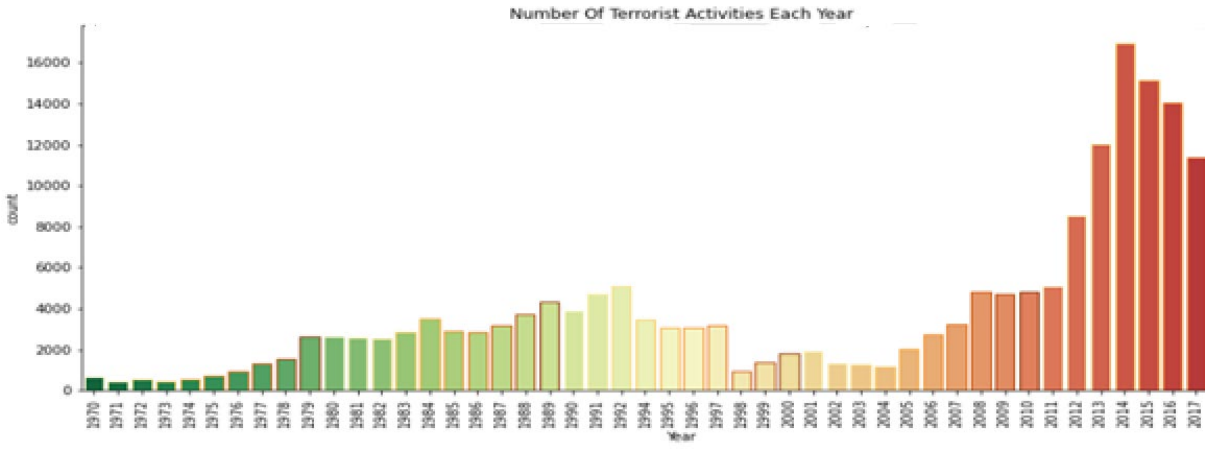

Figure 29. Numbers of terrorist activities of each year

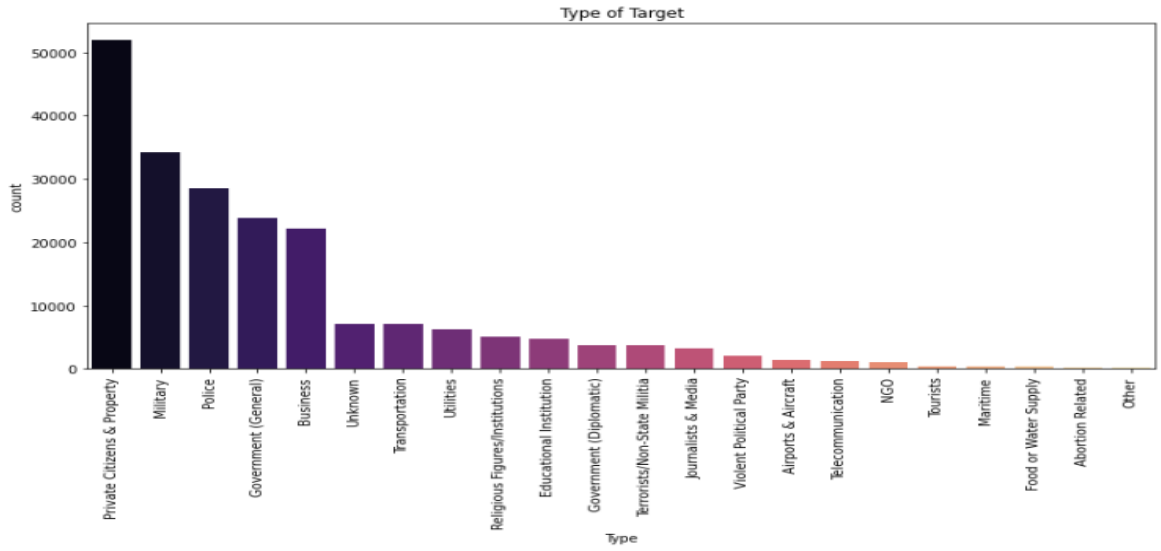

Figure 30. The most target type

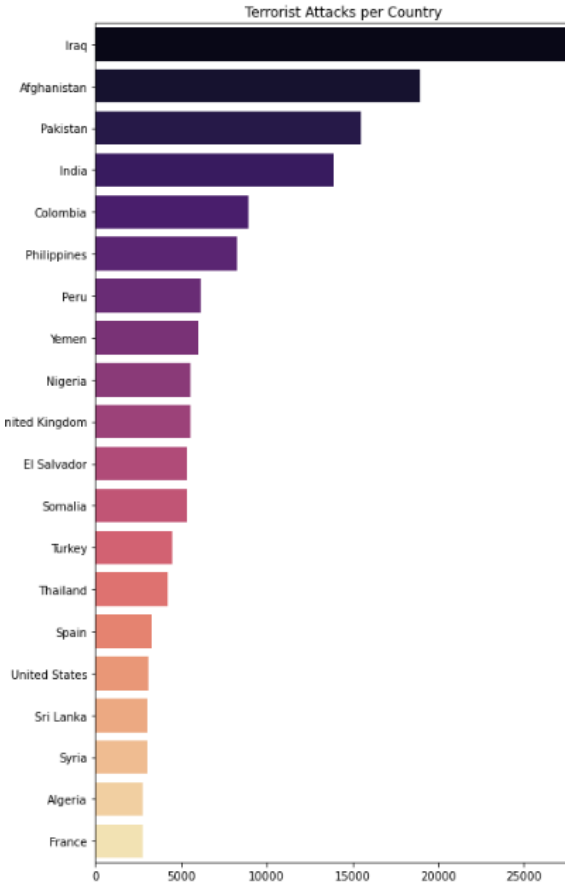

Figure 31. Terrorist attacks by Country

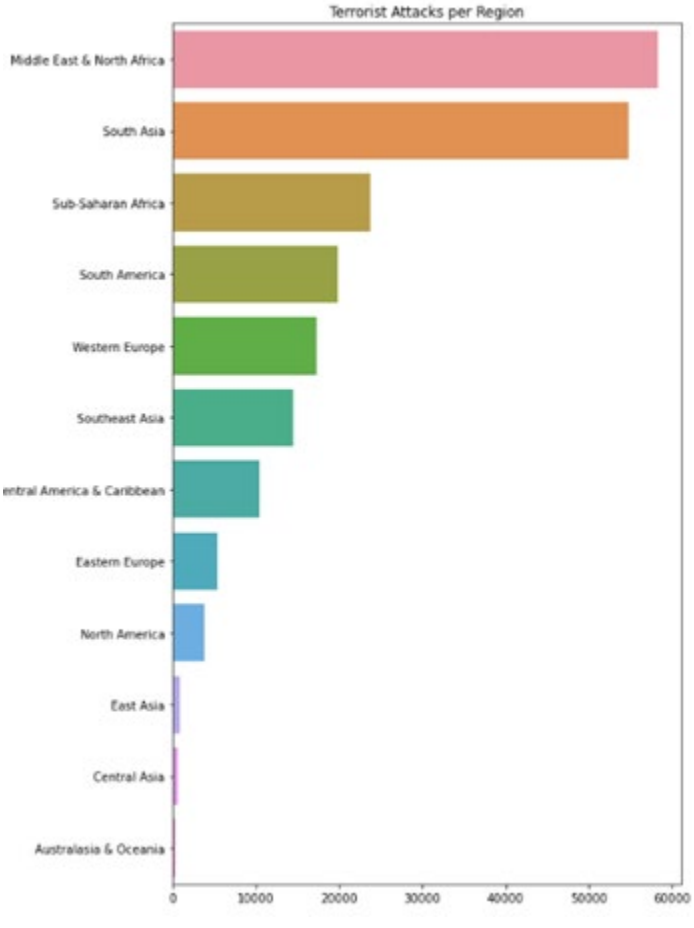

Figure 32. Terrorist attacks by regions

```
Iraq has suffered the maximum number of terror attacks of 27521
Andorra has suffered the minimum number of terror attacks of 1
```

Figure 33. The country suffers the max/min number of attacks

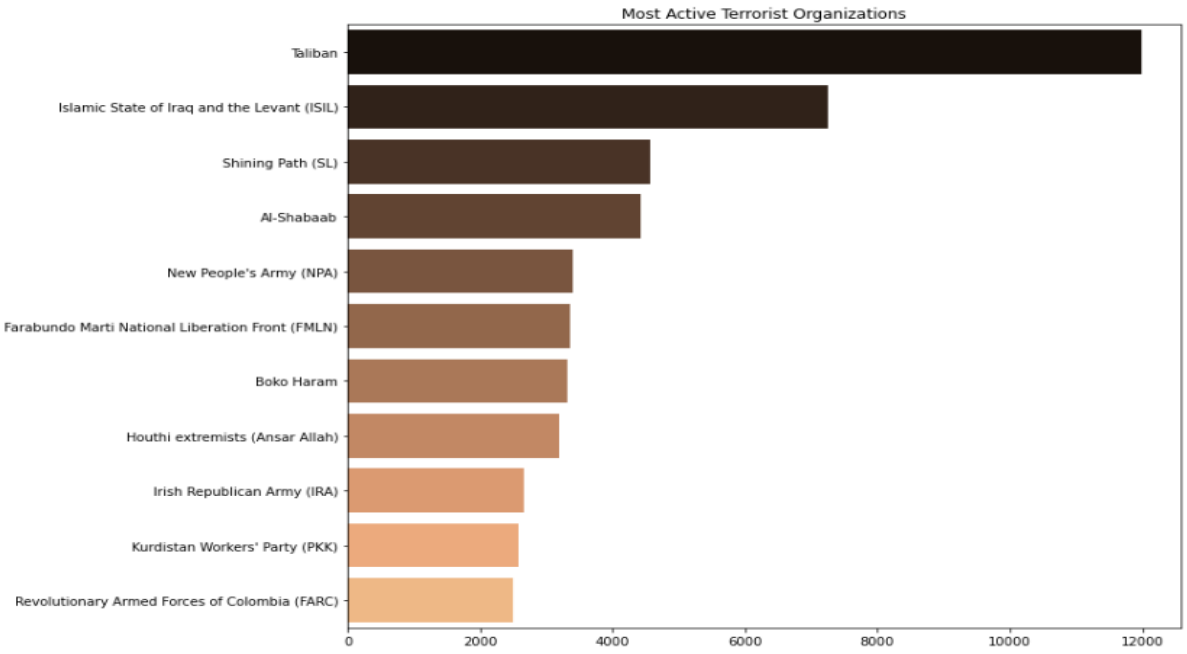

Figure 34. Most active terrorist Organisation

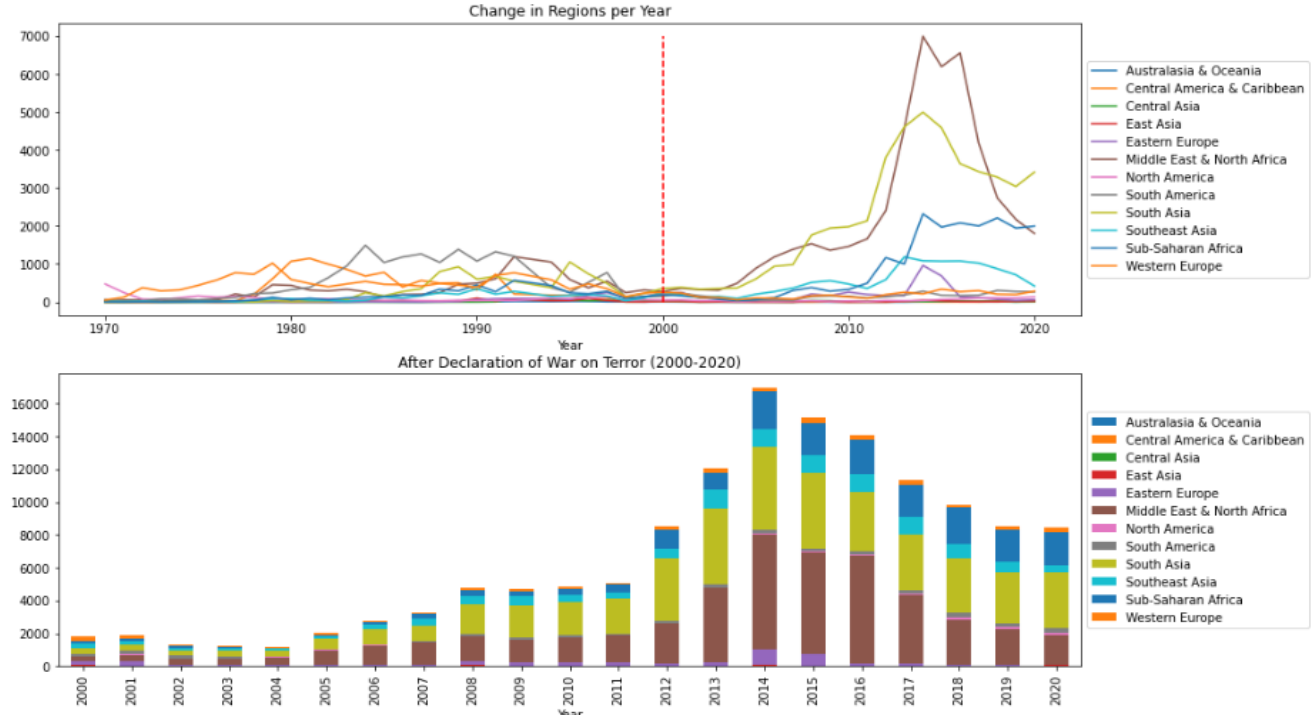

Figure 35. What are the changes after the declaration of war against terror in 2000

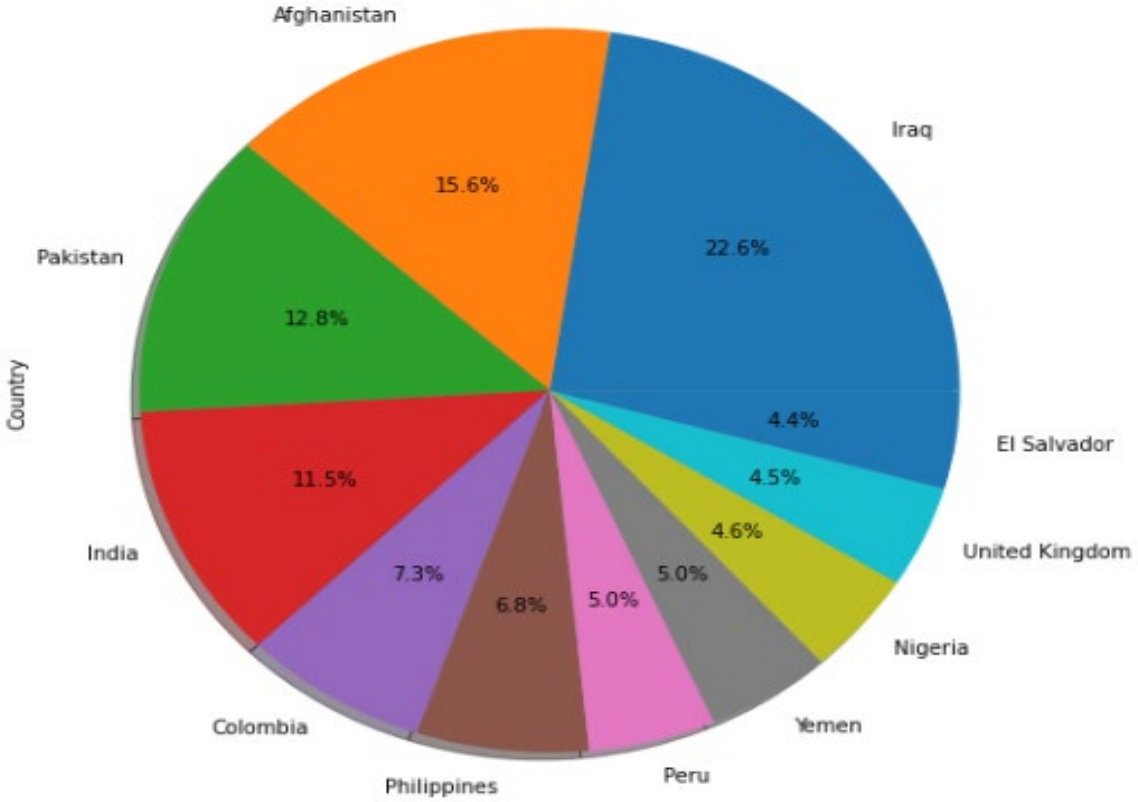

Figure 36. Terrorist attacks by Country

5.29 Observations

- The data consist of terrorist activities ranging from the year: 1970 to 2020
- The maximum number of people killed in an event was were:1700
- The maximum number of people wounded in an event was were:10878
- Max number of total casualties in an event was were:12263
- 651 attacks happened in 1970 & 8438 attacks happened in 2020
- The number of attacks from 1970 has increased by 92.0 % till 2020
- Iraq has the most terrorist attacks
- Since the Declaration of War on Terror in 2000, terrorist attack has increased
- Andorra has the minimum terrorist attack
- Terrorist Attacks per Region most terrorist attacks the Middle East & Africa
- Most Target type civilians and private property
- Most methods of Attack bombing explosion

6. Prediction

A Python predictive model predicts a particular product assurance based on prior data patterns. Fundamentally, by gathering and analysing historical data, you may train a pattern-detection model that can predict outcomes including future sales, disease spread, fraud, and many others.

	eventid	iyear	imonth	iday	approxdate	extended	resolution	country	country_txt	region	...	addnotes	scite1	scite2	scite3	dbsource	INT_LOG
0	197000000001	1970	7	2	NaN	0	NaN	58	Dominican Republic	2	...	NaN	NaN	NaN	NaN	PGIS	0
1	197000000002	1970	0	0	NaN	0	NaN	130	Mexico	1	...	NaN	NaN	NaN	NaN	PGIS	0
2	197001000001	1970	1	0	NaN	0	NaN	160	Philippines	5	...	NaN	NaN	NaN	NaN	PGIS	-9
3	197001000002	1970	1	0	NaN	0	NaN	78	Greece	8	...	NaN	NaN	NaN	NaN	PGIS	-9
4	197001000003	1970	1	0	NaN	0	NaN	101	Japan	4	...	NaN	NaN	NaN	NaN	PGIS	-9

Figure 37. Dataset

6.1 Futures pre-processing

Before Cleaning and processing, visualise the database and here we derived extrapolated, the most terrorism attacked country every 10 years. To get the result need it from the data, after exploration and analyse the context is driven to just 4 features eventid, iyear, country, country_txt.

	eventid	iyear	country	country_txt
0	197000000001	1970	58	Dominican Republic
1	197000000002	1970	130	Mexico
2	197001000001	1970	160	Philippines
3	197001000002	1970	78	Greece
4	197001000003	1970	101	Japan
...	...	...	...	...
209701	202012310015	2020	228	Yemen
209702	202012310016	2020	228	Yemen
209703	202012310017	2020	75	Germany
209704	202012310018	2020	4	Afghanistan
209705	202012310019	2020	33	Burkina Faso

209706 rows × 4 columns

Figure 38. Futures pre-processing

6.2 Analyse dataset by decades

Which is the most terrorism-affected country each decade

Use the original database. Generate the column decade (10 years), customs value from the column ‘iyear’ and stretches for every 10 years. after that represent the data for the ne

```
df1['decade'].value_counts()
```

```
2020s    114815
1980s     31156
1990s     28765
2000s     25057
1970s      9913
Name: decade, dtype: int64
```

Figure 39. Attacks by decades

Result description of the new dataset representation

	eventid	iyear	country	country_txt	decade
0	197000000001	1970	58	Dominican Republic	1970s
1	197000000002	1970	130	Mexico	1970s
2	197001000001	1970	160	Philippines	1970s
3	197001000002	1970	78	Greece	1970s
4	197001000003	1970	101	Japan	1970s
...	...	...	...	...	...
209701	202012310015	2020	228	Yemen	2020s
209702	202012310016	2020	228	Yemen	2020s
209703	202012310017	2020	75	Germany	2020s
209704	202012310018	2020	4	Afghanistan	2020s
209705	202012310019	2020	33	Burkina Faso	2020s

209706 rows × 5 columns

Figure 40. Target type assassinations by decade

country_txt	eventid	iyear	country	decade
Iraq	27521	27521	27521	27521
Afghanistan	18920	18920	18920	18920
Pakistan	15504	15504	15504	15504
India	13929	13929	13929	13929
Colombia	8915	8915	8915	8915
Philippines	8271	8271	8271	8271
Peru	6111	6111	6111	6111
Yemen	6027	6027	6027	6027
Nigeria	5550	5550	5550	5550
nited Kingdom	5513	5513	5513	5513
El Salvador	5320	5320	5320	5320
Somalia	5317	5317	5317	5317
Turkey	4485	4485	4485	4485

country_txt	eventid	iyear	country	decade
Iraq	27515	27515	27515	27515
Afghanistan	18916	18916	18916	18916
Pakistan	15487	15487	15487	15487
India	13905	13905	13905	13905
Colombia	8496	8496	8496	8496
Philippines	8147	8147	8147	8147
Peru	6092	6092	6092	6092
Yemen	6027	6027	6027	6027
Nigeria	5549	5549	5549	5549
Somalia	5316	5316	5316	5316
El Salvador	4875	4875	4875	4875
Thailand	4175	4175	4175	4175
Turkey	4001	4001	4001	4001

country_txt	eventid	iyear	country	decade
Iraq	27485	27485	27485	27485
Afghanistan	18894	18894	18894	18894
Pakistan	15291	15291	15291	15291
India	12677	12677	12677	12677
Philippines	7229	7229	7229	7229
Yemen	6027	6027	6027	6027
Colombia	5545	5545	5545	5545
Nigeria	5543	5543	5543	5543
Somalia	5299	5299	5299	5299
Thailand	4101	4101	4101	4101
Turkey	3661	3661	3661	3661
Syria	2855	2855	2855	2855
Algeria	2749	2749	2749	2749

country_txt	eventid	iyear	country	decade
Iraq	27360	27360	27360	27360
Afghanistan	18796	18796	18796	18796
Pakistan	13687	13687	13687	13687
India	10687	10687	10687	10687
Philippines	6283	6283	6283	6283
Yemen	5920	5920	5920	5920
Nigeria	5469	5469	5469	5469
Somalia	5148	5148	5148	5148
Thailand	3975	3975	3975	3975
Syria	2851	2851	2851	2851
Colombia	2712	2712	2712	2712
Libya	2491	2491	2491	2491
Egypt	2179	2179	2179	2179
Turkey	1996	1996	1996	1996
Russia	1914	1914	1914	1914
Ukraine	1792	1792	1792	1792

country_txt	eventid	iyear	country	decade
Iraq	22174	22174	22174	22174
Afghanistan	16845	16845	16845	16845
Pakistan	11707	11707	11707	11707
India	8315	8315	8315	8315
Yemen	5825	5825	5825	5825
Philippines	5262	5262	5262	5262
Nigeria	5212	5212	5212	5212
Somalia	4663	4663	4663	4663
Syria	2848	2848	2848	2848
Thailand	2747	2747	2747	2747
Libya	2489	2489	2489	2489
Egypt	2166	2166	2166	2166
Ukraine	1783	1783	1783	1783
Turkey	1732	1732	1732	1732
Colombia	1697	1697	1697	1697

country_txt	eventid	iyear	country	decade
Iraq	22174	22174	22174	22174
Afghanistan	16845	16845	16845	16845
Pakistan	11707	11707	11707	11707
India	8315	8315	8315	8315
Yemen	5825	5825	5825	5825
Philippines	5262	5262	5262	5262
Nigeria	5212	5212	5212	5212
Somalia	4663	4663	4663	4663
Syria	2848	2848	2848	2848
Thailand	2747	2747	2747	2747
Libya	2489	2489	2489	2489
Egypt	2166	2166	2166	2166
Ukraine	1783	1783	1783	1783

Figure 41. Assassination by decades (plain representation)

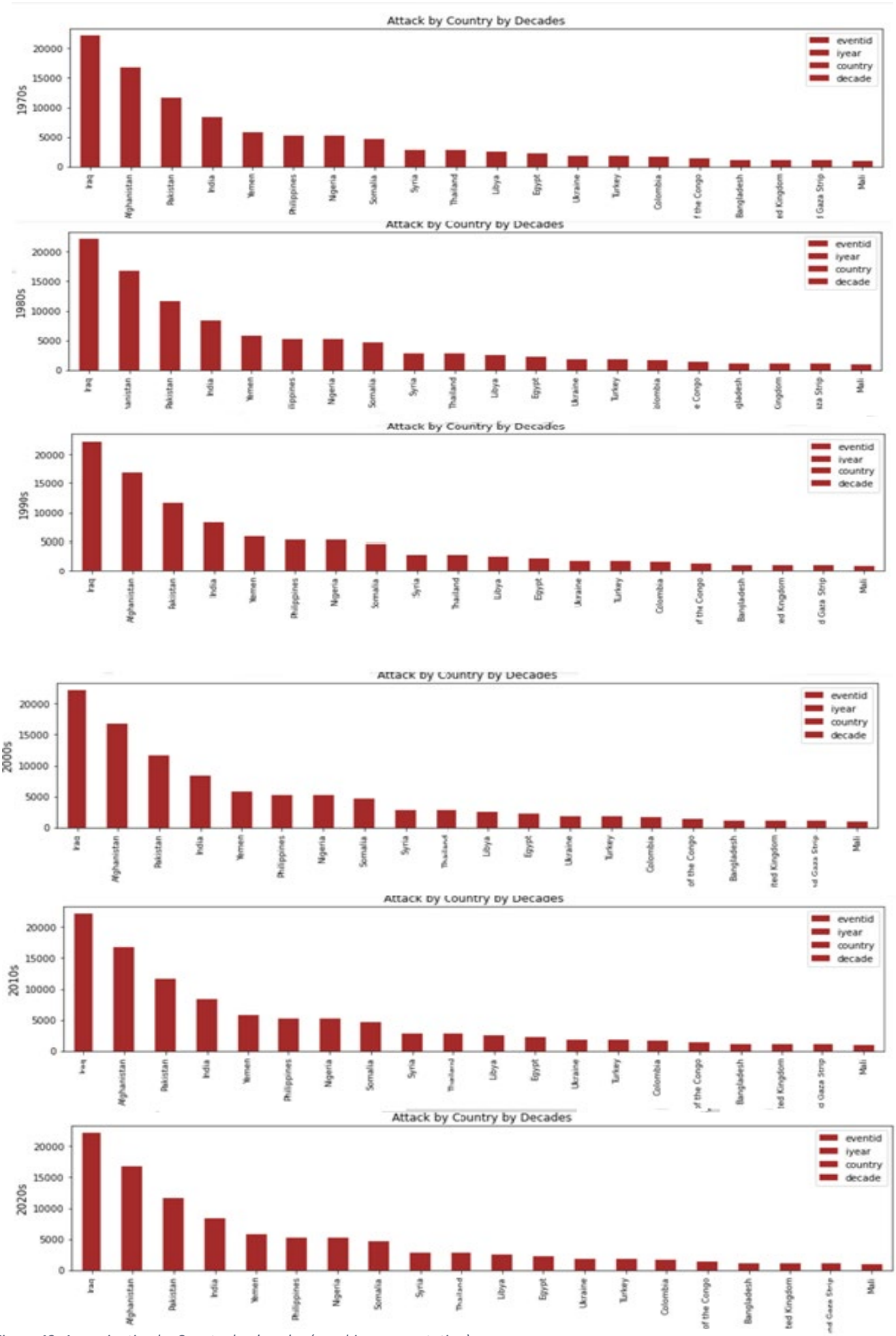

Figure 42. Assassination by Country by decades (graphic representation)

	eventid	attacktype1_txt	targtype1_txt	targsubtype1_txt	targtype1
0	197000000001	Assassination	Private Citizens & Property	Named Civilian	14
1	197000000002	Hostage Taking (Kidnapping)	Government (Diplomatic)	Diplomatic Personnel (outside of embassy, cons...	7
2	197001000001	Assassination	Journalists & Media	Radio Journalist/Staff/Facility	10
3	197001000002	Bombing/Explosion	Government (Diplomatic)	Embassy/Consulate	7
4	197001000003	Facility/Infrastructure Attack	Government (Diplomatic)	Embassy/Consulate	7
...	...	...	...	...	...
209701	202012310015	Bombing/Explosion	Private Citizens & Property	House/Apartment/Residence	14
209702	202012310016	Bombing/Explosion	Private Citizens & Property	House/Apartment/Residence	14
209703	202012310017	Facility/Infrastructure Attack	Military	Military Transportation/Vehicle (excluding con...	4
209704	202012310018	Armed Assault	Private Citizens & Property	Protester	14
209705	202012310019	Armed Assault	Military	Paramilitary	4

209706 rows × 5 columns

Figure 43. Filter pre-processing column

The result after cleaning all unwanted data caused by NAN and null

	eventid	attacktype1_txt	targtype1_txt	targsubtype1_txt	targtype1
0	197000000001	Assassination	Private Citizens & Property	Named Civilian	14
1	197000000002	Hostage Taking (Kidnapping)	Government (Diplomatic)	Diplomatic Personnel (outside of embassy, cons...	7
2	197001000001	Assassination	Journalists & Media	Radio Journalist/Staff/Facility	10
3	197001000002	Bombing/Explosion	Government (Diplomatic)	Embassy/Consulate	7
4	197001000003	Facility/Infrastructure Attack	Government (Diplomatic)	Embassy/Consulate	7
5	197001010002	Armed Assault	Police	Police Building (headquarters, station, school)	3
6	197001020001	Assassination	Police	Police Security Forces/Officers	3

Figure 44 a cleaning data

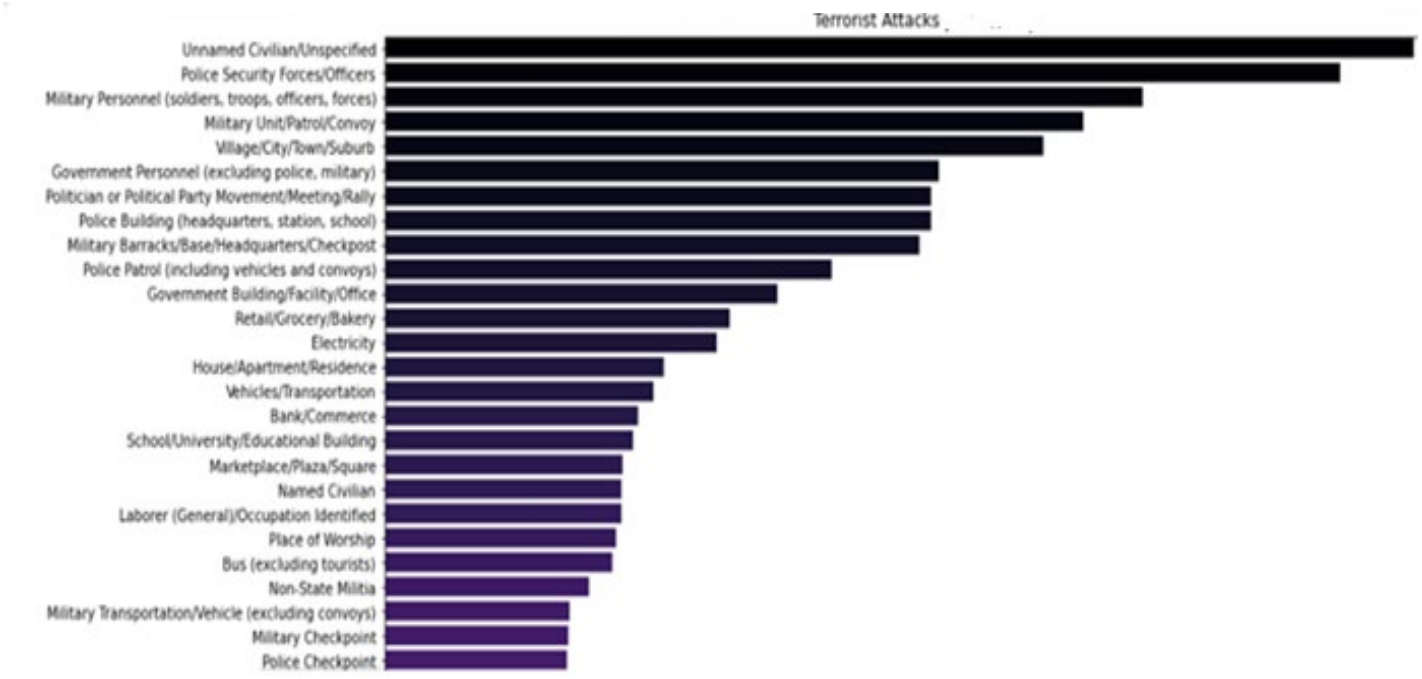

Figure 45. Target type

6.3 **Analyse the most active terrorist groups**

	eventid	gname	weapsubtype1_txt	nkill
0	197000000001	MANO-D	NaN	1.0
1	197000000002	23rd of September Communist League	NaN	0.0
2	197001000001	Unknown	NaN	1.0
3	197001000002	Unknown	Unknown Explosive Type	NaN
4	197001000003	Unknown	NaN	NaN
...	...	...	...	...
209701	202012310015	Houthi extremists (Ansar Allah)	Projectile (rockets, mortars, RPGs, etc.)	NaN
209702	202012310016	Houthi extremists (Ansar Allah)	Landmine	NaN
209703	202012310017	Left-wing extremists	Arson/Fire	0.0
209704	202012310018	Unknown	Unknown Gun Type	1.0
209705	202012310019	Unknown	Unknown Gun Type	5.0

209706 rows × 4 columns

Figure 46. Filter futures pre-processing

	eventid	gname	weapsubtype1_txt	nkill
5	197001010002	Black Nationalists	Unknown Gun Type	0.0
6	197001020001	Tupamaros (Uruguay)	Automatic or Semi-Automatic Rifle	0.0
8	197001020003	New Year's Gang	Molotov Cocktail/Petrol Bomb	0.0
9	197001030001	New Year's Gang	Gasoline or Alcohol	0.0
10	197001050001	Weather Underground, Weathermen	Unknown Explosive Type	0.0
...	...	...	...	...
209690	202012310003	Taliban	Landmine	1.0
209692	202012310005	The Resistance Front (TRF)	Unknown Gun Type	1.0
209695	202012310008	Patriotic Resistance Front in Ituri (FRPI)	Unknown Gun Type	0.0
209698	202012310012	Allied Democratic Forces (ADF)	Arson/Fire	25.0
209703	202012310017	Left-wing extremists	Arson/Fire	0.0

92715 rows × 4 columns

Figure 47. Drop the NAN, NULL and unknown from the dataset

Count Group has the greatest number of attacks

Data have too many entries, which is not upright for analysing. Focalised on data having more than one thousand to get the sum for all other values.

	weapsubtype1_txt	nkill
0	Unknown Gun Type	107893.0
1	Automatic or Semi-Automatic Rifle	80739.0
2	Vehicle	61970.0
3	Unknown Explosive Type	40898.0
4	Suicide (carried bodily by human being)	32381.0
5	Projectile (rockets, mortars, RPGs, etc.)	29163.0
6	Landmine	12064.0
7	Knife or Other Sharp Object	11570.0
8	Handgun	8184.0
9	Grenade	6365.0
10	Other Explosive Type	5609.0

Figure 48. Groups with the most attacks and the sum for the other value is 59275

6.4 Economical loss of each decade and its comparison based on the criticality of the attack.

Decade variable like we did in the point above. This is how our data looks after adding decades.

	eventid	iyear	country_txt	propvalue	crit1
0	197000000001	1970	Dominican Republic	NaN	1
1	197000000002	1970	Mexico	NaN	1
2	197001000001	1970	Philippines	NaN	1
3	197001000002	1970	Greece	NaN	1
4	197001000003	1970	Japan	NaN	1
...	...	...	...	...	...
209701	202012310015	2020	Yemen	-99.0	1
209702	202012310016	2020	Yemen	NaN	1
209703	202012310017	2020	Germany	-99.0	1
209704	202012310018	2020	Afghanistan	NaN	1
209705	202012310019	2020	Burkina Faso	NaN	1

209706 rows × 5 columns

Figure 49. Data filter

	eventid	iyear	country_txt	propvalue	crit1
7	197001020002	1970	United States	22500.0	1
8	197001020003	1970	United States	60000.0	1
10	197001050001	1970	United States	0.0	1
11	197001060001	1970	United States	305.0	1
14	197001090002	1970	United States	2000000.0	1
...	...	...	...	...	...
209697	202012310010	2020	Greece	-99.0	1
209699	202012310013	2020	Yemen	-99.0	1
209700	202012310014	2020	Nepal	-99.0	1
209701	202012310015	2020	Yemen	-99.0	1
209703	202012310017	2020	Germany	-99.0	1

48375 rows × 5 columns

Figure 50. Drop n/a

	eventid	iyear	country_txt	propvalue	crit1
0	197000000001	1970	Dominican Republic	NaN	1
1	197000000002	1970	Mexico	NaN	1
2	197001000001	1970	Philippines	NaN	1
3	197001000002	1970	Greece	NaN	1
4	197001000003	1970	Japan	NaN	1
...	...	...	...	...	...
209696	202012310009	2020	Iraq	NaN	1
209698	202012310012	2020	Democratic Republic of the Congo	NaN	1
209702	202012310016	2020	Yemen	NaN	1
209704	202012310018	2020	Afghanistan	NaN	1
209705	202012310019	2020	Burkina Faso	NaN	1

171789 rows × 5 columns

Figure 52 Result after drop the -99 from propvalue column

Economical loss value of private property

	iyear	propvalue
0	1992	2.806137e+09
1	1996	1.296775e+09
2	1982	7.482039e+08
3	1995	6.553832e+08
4	2001	3.666540e+08
5	2020	2.644970e+08
6	2016	2.118942e+08
7	1991	1.874643e+08
8	1989	1.646855e+08
9	1980	1.581276e+08
10	1977	1.463380e+08
11	2002	1.331200e+08
12	2014	1.272726e+08
13	1981	1.006995e+08
14	1987	9.389913e+07

Figure 51. economical loss on private property

6.5 Logistic Regression, Random Forest, Gaugasian to see the success of the terrorist attack

- The objective of logistic regression is to address categorization issues. In contrast to linear regression, which predicts a continuous result, categorical regression predicts discrete outcomes. In the simplest scenario, there are two results, known as binomial, such as determining whether a true or false

- Random Forest is a form of supervised machine learning technique that relies on ensemble learning. Supervised methods are a learning in which several types of algorithms or numerous instances of the same algorithm are combined to create a more accurate prediction model.

- The Gaussian Process is a method for machine learning. It may be used for regression and classification, among other tasks. As a Bayesian technique, Gaussian Process produces questionable predictions. For instance, it will forecast that the stock price for tomorrow will be $300, with a standard deviation of $50.

6.5.1 Analysing the Dataset

Determine the success rate of terrorist attacks in various countries. First, we create dummies to perform Random Forest Regression - Logistic Regression - GaussianNB.

	success	attacktype1	targtype1_txt	nkill
0	1	1	Private Citizens & Property	1.0
1	1	6	Government (Diplomatic)	0.0
2	1	1	Journalists & Media	1.0
3	1	3	Government (Diplomatic)	NaN
4	1	7	Government (Diplomatic)	NaN
...	...	...	...	...
209701	1	3	Private Citizens & Property	NaN
209702	1	3	Private Citizens & Property	NaN
209703	1	7	Military	0.0
209704	1	2	Private Citizens & Property	1.0
209705	1	2	Military	5.0

209706 rows × 4 columns

Figure 53 Select the Futures

	success	attacktype1	targtype1_txt	nkill
0	1	1	Private Citizens & Property	1.0
1	1	6	Government (Diplomatic)	0.0
2	1	1	Journalists & Media	1.0
3	1	3	Government (Diplomatic)	NaN
4	1	7	Government (Diplomatic)	NaN
...	...	...	...	...
209701	1	3	Private Citizens & Property	NaN
209702	1	3	Private Citizens & Property	NaN
209703	1	7	Military	0.0
209704	1	2	Private Citizens & Property	1.0
209705	1	2	Military	5.0

209706 rows × 4 columns

Figure 54. Drop the n/a from dataset

Logistic Regression Model

If there is a binary result, we apply the logistic regression modelling approach. For categorical variables, we must build dummy variables. Simply adding a number value to each category constitutes the creation of dummy variables. And then translating the category-containing rows into numerous columns. Use the following code and then execute df.head() to see the addition of columns to the right end of the table.

	success	attacktype1	nkill	targtype1_txt_Airports & Aircraft	targtype1_txt_Business	targtype1_txt_Educational Institution	targtype1_txt_Food or Water Supply	targtype1_txt_Government (Diplomatic)
0	1	1	1.0	0	0	0	0	0
1	1	6	0.0	0	0	0	0	1
2	1	1	1.0	0	0	0	0	0
5	1	2	0.0	0	0	0	0	0
6	0	1	0.0	0	0	0	0	0
...	...	...	...	...	...	...	...	...
209699	1	3	0.0	0	0	0	0	0
209700	1	7	0.0	0	0	0	0	0
209703	1	7	0.0	0	0	0	0	0
209704	1	2	1.0	0	0	0	0	0
209705	1	2	5.0	0	0	0	0	0

197179 rows × 24 columns

Figure 55. Build Dummy variables

Stats models is a Python library that enables users to investigate data, estimate statistical models, and conduct statistical tests. For various data types and estimators,

a comprehensive array of descriptive statistics, statistical tests, charting routines, and outcome statistics is accessible.

```
Optimization terminated successfully.
         Current function value: 0.152410
         Iterations 18
                           Logit Regression Results
==============================================================================
Dep. Variable:                success   No. Observations:                45866
Model:                          Logit   Df Residuals:                    45858
Method:                           MLE   Df Model:                            7
Date:                Wed, 07 Sep 2022   Pseudo R-squ.:                 0.03799
Time:                        03:51:51   Log-Likelihood:                -6990.5
converged:                       True   LL-Null:                       -7266.5
Covariance Type:            nonrobust   LLR p-value:                4.970e-115
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     12.0589      5.439      2.217      0.027       1.399      22.719
iyear         -0.0050      0.003     -1.851      0.064      -0.010       0.000
propvalue   3.802e-05   4.81e-06      7.906      0.000    2.86e-05    4.75e-05
nkill          0.1667      0.015     11.090      0.000       0.137       0.196
country       -0.0006      0.000     -2.475      0.013      -0.001      -0.000
crit1          0.7039      0.169      4.166      0.000       0.373       1.035
crit2          0.5663      0.315      1.799      0.072      -0.051       1.183
crit3         -0.3010      0.095     -3.153      0.002      -0.488      -0.114
==============================================================================
```

Figure 56. Data investigation

The measure of association between an exposure and an outcome.

```
Intercept    172628.416995
iyear             0.995031
propvalue         1.000038
nkill             1.181365
country           0.999434
crit1             2.021698
crit2             1.761811
crit3             0.740077
dtype: float64
```

Figure 57. Outcome measure

Logistic Regression Coefficients

Interpreting linear regression coefficients is straightforward, and you would verbally describe the coefficients like this:

“For every one-unit increase in [X variable], the [y variable] increases by [coefficient] when all other variables are held constant.”

```
 0.05466816  0.15216226 -0.40791298  0.62588632  0.3696828   0.18255818
-0.27176298 -0.29960908  0.11593029 -0.19189746 -0.10035262  0.74397506
 0.20085801  0.25064558  0.51827761  0.40124402  0.90088094  0.31040065
 0.49659483  0.15656221 -2.50047069  0.92821759 -0.34351817]] [1.61199979]
```

Figure 58. Coefficient

Classification report

A Classification report is used to evaluate the accuracy of a classification algorithm's predictions. How many forecasts were accurate and how many were inaccurate? As seen here, True Positives, False Positives, True Negatives, and False Negatives are used to forecast the metrics of a categorization report.

```
'              precision    recall  f1-score   support\n\n           0       0.66      0.19      0.30     11799\n           1
0.90      0.99      0.94     86791\n\n    accuracy                           0.89     98590\n   macro avg       0.78      0.59
0.62     98590\nweighted avg       0.87      0.89      0.86     98590\n'
```

Figure 59. Classification report

Confusion matrix

The confusion matrix is a table used to characterise the performance of a classification method. A confusion matrix is a graphical representation and summary of the performance of a classification system.

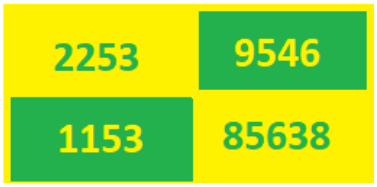

Figure 60. Confusion matrix

Models score

Random Forest Regressor	GaussianNB	Logistic Regression
0.30411296841564117	0.7059742367380059	0.8914798661121818

Table 1

7. Conclusion

This project in analysing global terrorism around the globe assisted in learning effective data cleansing techniques and how to separate and filter data. It taught me how to identify and separate the essential data necessary for analysis from most raw data.

The best model best attained a precision is **Logistic Regression 0.89.**

Even after the regularisation of the models offered a better grasp of data manipulation through many functions and techniques that produced the intended outcome.

We employed statistical approaches to data, such as logistic regression, which assisted us in analysing the relationship between variables, such as how a change in one variable would affect and alter several data parameters.

The data selection process presented several significant obstacles. Two days were devoted to finding the optimal method for selecting appropriate data for questions. Losing vital data when doing fundamental data cleansing processes.

Separately sanitised the dataset for each question with just the needed data. In addition, build columns when appropriate. Sometimes operate only with 45 000 data entries out of 120,000+ entries. Segregation was also a significant element since the volume of data inputs made it difficult at times to analyse the displayed data. Another difficulty is that Anaconda is unstable; the computer was lost for more than a week and had to be formatted twice.

- Using project management to split the project into small projects into modules, days, and hours to study and in just three to six hours a day.
- Conduct a study on Data Analyse applications.

Conducted a comprehensive study on how to manage the project and on how analyser data analysts carry out their day-to-day activities. A study was also conducted on a similar project available in the market. Studied some of the applications available from leading software vendors. World Wide Web was used as the main medium of study. The study contributed a lot to the research by giving an overall picture of the area of the research.

- Analyse the requirements of project applications.

The requirement-gathering phase of the project was carried out by reading literature available on the World Wide Web. Some of the requirements were also identified from the study conducted on the similar tools available. The requirements gathered contributed a lot towards preparing requirement specifications.

- A study on the technology

Studies the materials on the technologies, available for implementation. Explored alternative technologies available and then decided on the best that suits. The study was documented to be attached in the final documentation.

Also carried out a study on the tools that can be used during the implementation of the proposed solution. After considering the alternative tools available and choosing the appropriate one,

- Prepare the specification of the recommended solution.
- Present all observations in the form of a report.

One of the primary objectives was to document the form of observations, findings, and lessons in the form of a report. This was achieved by the thesis for the project. The thesis, evaluation observations, comments, and experiences

The research paved the way to apply the knowledge gained during the degree programme. Even though virtually all the modules learned contributed to the success of the project most prominent ones are listed below.

Module	Use in Context of the Project
Research methods and proposal	Literature survey and evaluation
Data analytics and data management	Managing data analysing
Anaconda software systems engineering	Python concept in anaconda
System analysis	System development approach
Jupiter Notebook	Database related concepts
Perspective of information technology	Python related concepts

Table 2

Completing such an academic project is by no means an easy task given the time constraints. However, the crucial factor is that the entire duration was a learning curve. The exposure was theoretical. Even though the outcome of the project was the most valuable outcome was the knowledge and experiences gathered during this period.

The importance of Planning and Scheduling is the most important lesson learned by the author. The activity schedule prepared at the inception stage of the project was used to monitor the progress of the research. Some of the activities were revised due to some new ones arriving. Those new ones were not identified during the research start-up. This was a lesson learnt a hard way by the writer.

Developed project management skills even though the solution was not developed. Project management includes time management, task scheduling and allocation. A study on Python provided the author with a sound understanding of the concepts behind Python. Theoretical knowledge gained was vital. Many lessons both academic and non-academic were learnt during the project which will help the writer immensely in the future.

- **System uses for project**

OS Name	Microsoft Windows 11 Pro
Version	10.0.22000 Build 22000
Other OS Description	Not Available
OS Manufacturer	Microsoft Corporation
System Name	LAUR
System Manufacturer	ASUSTeK COMPUTER INC.
System Model	VivoBook_ASUSLaptop X509JA_A509JA
System Type	x64-based PC
System SKU	
Processor	Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz, 1501 Mhz, 4 Core(s), 8 Logica.
BIOS Version/Date	American Megatrends Inc. X509JA.310, 12/08/2021
SMBIOS Version	3.2
Embedded Controller Version	255.255
BIOS Mode	UEFI
BaseBoard Manufacturer	ASUSTeK COMPUTER INC.
BaseBoard Product	X509JA
BaseBoard Version	1.0
Platform Role	Mobile

Reference

Bigoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence. 2019;1(1):20–23. DOI: 10.1038/s42256-018-0004-1. - DOI

Bigo, D., Carrera, S., Hernanz, N., Jeandesboz, J., Parkin, J., Ragazzi, F., & Scherrer, A. (2013). Mass surveillance of personal data by EU member states and its compatibility with EU law. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2360473.

Bird SJ. Security and privacy: Why privacy matters. Science and Engineering Ethics. 2013;19(3):669–671. DOI: 10.1007/s11948-013-9458-z. - DOI - PubMed

Borowiec, S. (2016). AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol. Guardian. Retrieved from https://www.theguardian.com/technology/2016/mar/15/googles-alphago-seals...

Brayne S. Big data surveillance: The case of policing. American Sociological Review. 2017;82(5):977–1008. DOI: 10.1177/0003122417725865. - DOI

Calude CS, Longo G. The deluge of spurious correlations in big data. Foundations of Science. 2017;22(3):595–612. doi: 10.1007/s10699-016-9489-4. - DOI

Camacho-Collados M, Liberatore F. A decision support system for predictive police patrolling. Decision Support Systems. 2015;75:25–37. doi: 10.1016/J.DSS.2015.04.012. - DOI

de Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD. Unique in the crowd: The privacy bounds of human mobility. Scientific Reports. 2013;3(1):1376. doi: 10.1038/srep01376. - DOI - PMC - PubMed

De Montjoye YA, Radaelli L, Singh VK, Pentland AS. Unique in the shopping mall: On the reidentifiability of credit card metadata. Science. 2015 doi: 10.1126/science.1256297. - DOI - PubMed

Dunson DB. Statistics in the big data era: Failures of the machine. Statistics & Probability Letters. 2018;136:4–9. doi: 10.1016/J.SPL.2018.02.028. - DOI

Feigenbaum, J., & Koenig, J. (2014). On the feasibility of a technological response to the surveillance morass. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 10.1007/978-3-319-12400-1_23.

Jonas J, Harper J. Effective counterterrorism and the limited role of predictive data mining. Policy Analysis. 2006;584:1–12.

Kift Paula, Nissenbaum Helen. Metadata in context-an ontological and normative analysis of the NSA's bulk telephony metadata collection program. ISJLP. 2016;13:333.

L'Heureux A, Grolinger K, Elyamany HF, Capretz MAM. Machine learning with big data: Challenges and approaches. IEEE Access. 2017;5:7776–7797. doi: 10.1109/ACCESS.2017.2696365. - DOI

Landau S. Making sense from snowden: What is significant in the NSA surveillance revelations. IEEE Security and Privacy. 2013 doi: 10.1109/MSP.2013.90. - DOI

Lindekilde L, O'Connor F, Schuurman B. Radicalization patterns and modes of attack planning and preparation among lone-actor terrorists: An exploratory analysis. Behavioral Sciences of Terrorism and Political Aggression. 2019;11(2):113–133. doi: 10.1080/19434472.2017.1407814. - DOI

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., & Stumpe, M. C. (2017). Detecting cancer metastases on gigapixel pathology images. Retrieved from http://arxiv.org/abs/1703.02442.

Matijosaitiene I, McDowald A, Juneja V. Predicting safe parking spaces: A machine learning approach to geospatial urban and crime data. Sustainability. 2019;11(10):2848. doi: 10.3390/su11102848. - DOI

Mayer J, Mutchler P, Mitchell JC. Evaluating the privacy properties of telephone metadata. Proceedings of the National Academy of Sciences of the United States of America. 2016 doi: 10.1073/pnas.1508081113. - DOI - PMC - PubMed

Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. Proceedings—IEEE Symposium on Security and Privacy. 2008 doi: 10.1109/SP.2008.33. - DOI

National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2018). Global Terrorism Database. Retrieved from http://www.start.umd.edu/gtd.

Naughton, J. (2013). NSA surveillance: Do not underestimate the extraordinary power of metadata. Guardian. Retrieved from https://www.theguardian.com/technology/2013/jun/21/nsa-surveillance-meta....

Schneier, B. (2015). NSA doesn't need to spy on your calls to learn your secrets. Wired. Retrieved from https://www.wired.com/2015/03/data-and-goliath-nsa-metadata-spying-your-....

Sirseloudi MP. How to predict the unpredictable: On the early detection of terrorist campaigns. Defense & Security Analysis. 2005;21(4):369–386. doi: 10.1080/14751790520003454 21. - DOI

Soghoian C. Insecure flight: Broken boarding passes and ineffective terrorist watch lists. Policies and research in identity management. Boston, MA: Springer; 2008. pp. 5–21.

Van den Hoven J, Lokhorst G-J, Van de Poel I. Engineering and the problem of moral overload. Science and Engineering Ethics. 2012;18(1):143–155. doi: 10.1007/s11948-011-9277-z. - DOI - PMC - PubMed