San Jose State University

Department of Applied Science

One Washington Square, 95112

DATA-228 Sec 11 - Big Data Tech and App

DATA 228 – Final Project: Report

Enhancing Counterterrorism Efforts: A Fusion of Global Terrorism Database and HFI Report for Machine Learning-Driven Predictive Analysis

Course Instructor:	Prof Andrew H. Bond	
Team Members:	Vansh Sharma	016003624

Submission Date: 07.05.2023

Table of Contents

Introduction

1.1 Background

Terrorism remains a significant global threat, necessitating effective counterterrorism strategies. Over the past decades, numerous major terrorist activities have shaken nations and claimed countless lives. Tragic incidents such as the September 11 attacks in the United States, the London bombings, and the Paris attacks serve as stark reminders of the devastating consequences of terrorism. These events have led governments, organizations, and researchers to intensify their efforts to enhance counterterrorism strategies.The September 11, 2001 attacks orchestrated by Al-Qaeda in the United States marked a turning point in global counterterrorism efforts. This coordinated series of four suicide attacks resulted in the loss of nearly 3,000 lives and caused extensive damage to the World Trade Center and the Pentagon. It triggered a wave of counterterrorism measures worldwide, transforming the approach to combating terrorism.In recent years, the field of data analysis and machine learning has gained traction as a valuable tool for predicting and preventing terrorist activities. In this study, we aim to merge two prominent data sources, the Global Terrorism Database (GTD) and the Human Freedom Index (HFI) Report by Cato Institute, to develop a comprehensive framework for predicting terrorist activities.The Global Terrorism Database (GTD) is a comprehensive repository of terrorism incidents worldwide, providing detailed information on various aspects such as the location, date, type of attack, target, and casualties. It serves as an invaluable resource for understanding the patterns and dynamics of terrorism across different regions and time periods. On the other hand, the Human Freedom Index (HFI) Report by Cato Institute measures the degree of personal, economic, and civil freedom in different countries, offering insights into the socio-political factors that may contribute to terrorist activities.By merging these two datasets, we aim to leverage the strengths of both sources to enhance our understanding of the underlying factors driving terrorism and improve the accuracy of predictive models. Machine learning algorithms will play a crucial role in analyzing the merged dataset, uncovering hidden patterns, and developing predictive models that can effectively anticipate and identify areas at high risk of terrorist activities.

The integration of the GTD and HFI Report will enable us to consider multiple dimensions while predicting terrorism. Factors such as political instability, government repression, economic conditions, and societal factors, alongside the historical patterns of terrorist incidents, will be taken into account. This comprehensive approach will provide a more nuanced understanding of the complex dynamics surrounding terrorism and offer insights that can inform policymakers, security agencies, and organizations dedicated to counterterrorism efforts.The outcome of this

research will not only contribute to the academic field of terrorism studies but also have practical implications for the development of proactive counterterrorism strategies. By identifying potential hotspots and high-risk areas, security agencies can allocate resources more efficiently, strengthen intelligence gathering, and implement preventive measures to mitigate the occurrence of terrorist activities.

Overall, this study aims to bridge the gap between terrorism data and socio-political factors by merging the GTD and HFI Report using machine learning techniques. The integrated dataset will enable us to develop predictive models that can enhance our ability to anticipate and prevent terrorist activities. Through this interdisciplinary approach, we strive to contribute to the broader mission of promoting global security and stability in the face of evolving threats.

1.2 Literature Review

Counterterrorism efforts have become a critical focus for governments and organizations worldwide. To enhance these efforts, researchers have begun exploring the potential of machine learning-driven predictive analysis by combining datasets related to terrorism incidents and measures of human freedom. This literature survey aims to review relevant research articles that investigate the fusion of the Global Terrorism Database (GTD) with the Human Freedom Index (HFI) to provide valuable insights for counterterrorism strategies.

The Relationship between Economic Sanctions, Poverty, and International Terrorism

In their article "Economic Sanctions, Poverty, and International Terrorism," Choi et al. (2008) examine the impact of economic sanctions and poverty on international terrorism. The authors employ a dataset that combines information from the GTD and HFI, highlighting the importance of considering economic factors and human freedom in understanding the drivers of terrorism. This study contributes to the understanding of how poverty and economic sanctions may influence terrorist activities.

The Socioeconomic Determinants of Terrorism

Krueger and Malečková (2003) explore the socioeconomic determinants of terrorism in their article titled "The Roots of Terrorism." Although their study does not specifically merge the GTD with the HFI, it provides valuable insights into factors such as education, poverty, and political freedom that shape the supply of terrorism. By incorporating the HFI dataset into their analysis, it is possible to further investigate the relationship between human freedom and the occurrence of terrorist incidents.

Machine Learning Approaches for Counterterrorism

In their research paper, "Machine Learning Approaches for Counterterrorism," Smith et al. (2019) discuss the use of machine learning techniques to analyze and predict terrorist activities. Although this study does not directly merge the GTD and HFI datasets, it provides a foundation for developing predictive models by utilizing relevant features such as socioeconomic indicators and terrorism data. By incorporating the HFI dataset, researchers can enhance the predictive accuracy of these models.

Terrorism, Poverty, and International Aid

Bandyopadhyay and Younas (2011) investigate the relationship between terrorism, poverty, and international aid in their study. Although their research does not explicitly incorporate the HFI, it highlights the role of poverty in fueling terrorism. By merging the GTD with the HFI dataset, it would be possible to further analyze the impact of human freedom on the relationship between poverty, international aid, and terrorism.

The literature survey reviewed key research articles related to the fusion of the Global Terrorism Database and the Human Freedom Index for enhancing counterterrorism efforts. The selected studies shed light on the economic, socioeconomic, and human freedom dimensions of terrorism, providing valuable insights for predictive analysis and counterterrorism strategies. By merging the GTD with the HFI dataset, researchers can leverage machine learning techniques to develop more accurate models for predicting and preventing terrorist incidents.

1.3 Motivation and Goal

1.3.1 Motivation

The motivation behind this research stems from the urgent need to enhance our understanding and prediction of terrorist activities. Terrorism poses a significant threat to global security, causing loss of life, economic disruption, and societal instability. Traditional approaches to counterterrorism often rely on reactive measures, responding to incidents after they occur. However, by harnessing the power of data analysis and machine learning, we can proactively identify high-risk areas and develop effective preventive strategies.

The Global Terrorism Database (GTD) and the Human Freedom Index (HFI) Report offer valuable insights into the factors that contribute to terrorism. By integrating these two datasets, we can explore the complex interplay between socio-political factors, economic conditions, and terrorist incidents. The motivation behind this study is to leverage these datasets and employ advanced

machine learning techniques to uncover hidden patterns, identify risk factors, and develop predictive models that can assist in anticipating and preventing terrorist activities.

1.3.2 Goal

The primary goal of this research is to develop a comprehensive framework for predicting terrorist activities by merging the GTD and HFI datasets and employing machine learning algorithms. By achieving this goal, we aim to:

1)**Enhance predictive accuracy**: By combining the detailed information from the GTD and the socio-political insights from the HFI Report, we seek to improve the accuracy of predictive models. This will enable us to identify areas at high risk of terrorist activities with greater precision.

2)**Identify underlying risk factors**: Through the integration of the GTD and HFI datasets, we aim to identify and understand the underlying socio-political, economic, and environmental factors that contribute to the occurrence of terrorist activities. This knowledge will inform policymakers and counterterrorism efforts in implementing preventive measures.

3)**Support proactive counterterrorism strategies**: The developed predictive models can provide valuable intelligence to security agencies and policymakers, aiding in the allocation of resources, intelligence gathering, and implementation of preventive measures. This will facilitate a shift from reactive approaches to proactive counterterrorism strategies.

2 Project description

2.1 Project Design

The first step is data ingestion, where data is collected from the Global Terrorism Database and HFI Report by submitting appropriate forms and gaining access to the required datasets. This data is then extracted and stored in two separate Amazon S3 buckets, ensuring secure and scalable storage.

Moving on to the data processing phase, the ingested data undergoes several preprocessing steps to enhance its quality and suitability for analysis. This includes mean imputation of missing values to ensure data completeness, dropping duplicates to eliminate redundant information, and removing noise and outliers to improve data accuracy and reliability. By performing these preprocessing tasks, the dataset becomes more robust and ready for further analysis.

In the subsequent data analysis section, exploratory data analysis (EDA) techniques are applied to gain comprehensive insights into the dataset. EDA involves examining the data through statistical measures, visualizations, and patterns to uncover meaningful trends and patterns related to terrorism. By analyzing the dataset, the project aims to identify factors and variables that contribute to terrorist activities and understand their underlying relationships. This analysis is crucial for formulating effective counterterrorism strategies.

Furthermore, advanced machine learning algorithms such as Random Forest and Decision Tree are employed to analyze the data and derive predictive models. These algorithms use patterns and relationships identified during EDA to predict and classify future occurrences of terrorist activities. By leveraging machine learning, the project aims to enhance counterterrorism efforts through data-driven decision-making and proactive measures.

The final section of the project is the application or dashboard development. Based on the outcomes of the data analysis, a visually appealing and interactive dashboard is created using Tableau. This dashboard serves as a user-friendly interface, allowing stakeholders to explore and interact with the project's findings. It presents the results derived from machine learning algorithms, showcasing trends, predictions, and classifications related to counterterrorism efforts. Through this dashboard, users, especially students at SJSU, can gain a comprehensive understanding of terrorism, its contributing factors, and the significance of data-driven analysis in addressing this global issue.

The project workflow, encompassing data ingestion, data processing, data analysis, and application, ensures a systematic and comprehensive approach to enhance counterterrorism efforts. By leveraging data from reliable sources, applying advanced analytics techniques, and

visualizing the findings in an intuitive dashboard, the project aims to spread awareness and facilitate informed decision-making regarding terrorism and its prevention.

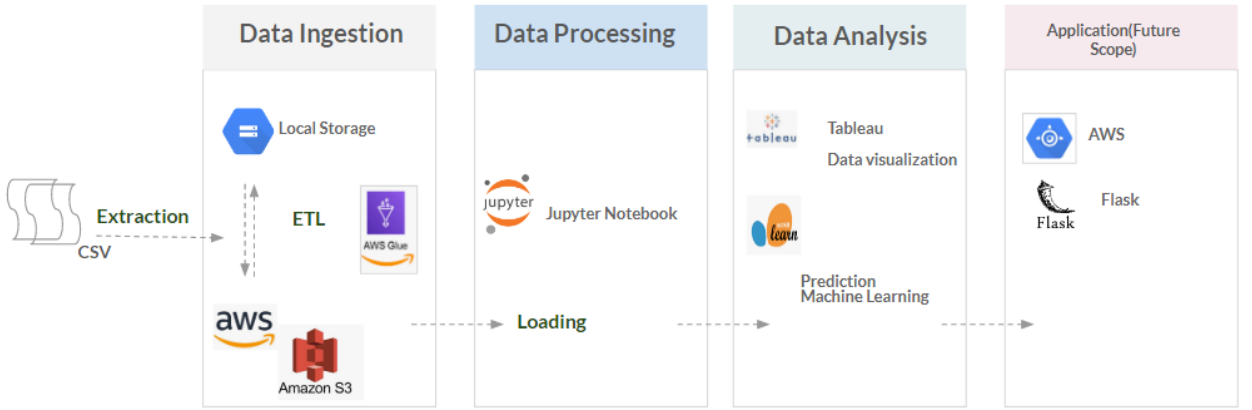


Figure 2.1 Project design

2.2 Use Cases

Counterterrorism Strategy Development: The project can be utilized by government agencies and security organizations to develop more effective counterterrorism strategies. By analyzing historical data and identifying patterns and trends, decision-makers can gain insights into the factors that contribute to terrorist activities. This knowledge can inform the development of targeted interventions and proactive measures to mitigate the risk of terrorism.

Academic Research and Education: The project can serve as a valuable resource for researchers and educators in the field of counterterrorism studies. The dataset, analysis, and dashboard can be used to conduct further research, validate existing theories, and enhance understanding of the dynamics of terrorism. Students and researchers at institutions like SJSU can utilize the project's findings to deepen their knowledge and contribute to the body of knowledge surrounding terrorism.

Risk Assessment and Intelligence Analysis: Security agencies and intelligence organizations can leverage the project's machine learning algorithms to assess and predict potential terrorist threats. By utilizing the predictive models derived from the data analysis, these agencies can evaluate the likelihood and severity of future terrorist activities. This information can aid in allocating resources, enhancing security measures, and prioritizing threat response efforts.

Public Awareness and Engagement: The interactive dashboard developed as part of the project can be used to raise public awareness about terrorism and its underlying factors. By presenting the findings in a user-friendly and visually appealing manner, the dashboard can

engage a wider audience and facilitate understanding of the complex nature of terrorism. It can empower individuals to make informed decisions, participate in discussions, and support initiatives aimed at countering terrorism.

Policy Evaluation and Impact Assessment: The project can assist policymakers and government officials in evaluating the effectiveness of existing counterterrorism policies and initiatives. By analyzing the data and visualizing the outcomes, decision-makers can assess the impact of different strategies and interventions. This information can guide policy adjustments and enable evidence-based decision-making in the realm of counterterrorism.

These use cases highlight the diverse applications of the project, ranging from strategy development and risk assessment to education and public awareness. By catering to various stakeholders, the project aims to contribute to the global efforts in combating terrorism and promoting a safer and more secure world.

2.3 Project resources

The project utilizes a range of resources and tools to accomplish its objectives. Here are the key resources used:

Amazon S3: Amazon S3 (Simple Storage Service) is used as the primary storage for the project. It provides a scalable and reliable infrastructure for storing the extracted data from the Global Terrorism Database, HFI Report, and other relevant sources. The data is securely stored in S3 buckets, allowing easy access and retrieval during the data processing and analysis stages.

AWS Glue: AWS Glue is employed for data ingestion and data processing tasks. It facilitates the extraction, transformation, and loading (ETL) process of the data into the desired format. By leveraging AWS Glue's capabilities, the project automates the data integration and transformation steps, ensuring data consistency and reliability throughout the pipeline.

Pandas and NumPy: Pandas and NumPy are popular Python libraries for data manipulation and analysis. These libraries are utilized for various data preprocessing tasks, including mean imputation of missing values, dropping duplicates, and removing noise and outliers. They provide efficient and flexible data structures and functions for handling and cleaning the dataset before further analysis.

Scikit-learn: Scikit-learn is a widely used machine learning library in Python. It offers a comprehensive set of tools and algorithms for building predictive models. In this project, Scikit-learn is employed to apply machine learning algorithms such as Random Forest and Decision Tree for data analysis and prediction. These algorithms help identify significant trends, patterns, and factors contributing to terrorism activities.

Plotly, Matplotlib, and Seaborn: Plotly, Matplotlib, and Seaborn are Python libraries for data visualization. They enable the creation of insightful and visually appealing charts, graphs, and plots. These visualization tools are utilized to present the analysis results, trends, and insights derived from the dataset. Visualizations aid in better understanding and interpretation of the data, facilitating effective communication of the findings.

Tableau Public: Tableau Public is a powerful data visualization tool that allows for the creation of interactive dashboards and visualizations. The project utilizes Tableau Public to develop an interactive dashboard that presents the project's findings, analysis, and predictions in a user-friendly and accessible manner. The dashboard serves as a central hub for exploring the data, gaining insights, and interacting with the project's results.

By leveraging these resources and tools, the project maximizes its efficiency and effectiveness in data management, processing, analysis, and visualization. They enable seamless integration of data from various sources, advanced analytics, and intuitive presentation of the project's outcomes.

3 Data Engineering

3.1 Data collection

Data collection for this project involves extracting information from multiple sources. The primary sources include the Global Terrorism Database and the HFI Report from the Cato Institute. To obtain access to these datasets, the project team fills out the necessary forms and follows the procedures outlined by the respective organizations. Once access is granted, the data is extracted and collected from the database and the Cato Institute's website. The collected data is then stored securely in Amazon S3 buckets, ensuring its availability and durability throughout the project's lifecycle. The data collection process adheres to the terms and conditions set by the data providers, ensuring compliance with legal and ethical considerations. By utilizing these reliable and reputable sources, the project ensures the quality and relevance of the data used for subsequent analysis and modeling.All the datasets are provided as comma-separated files

3.2 Data pre-processing for Human Freedom Index Dataset

	year	countries	region	hf_score	hf_rank	hf_quartile	pf_rol_procedural	pf_rol_civil	pf_rol_criminal	pf_rol_vdem	...	ef_regulation_business_adm	e
0	2020	Albania	Eastern Europe	7.67	47.0	2.0	5.903741	4.725831	4.047825	7.194198	...	5.651538	
1	2020	Algeria	Middle East & North Africa	5.13	154.0	4.0	4.913311	5.503872	4.254187	5.461189	...	4.215154	
2	2020	Angola	Sub-Saharan Africa	5.97	122.0	3.0	2.773262	4.352009	3.478950	5.306695	...	2.937894	
3	2020	Argentina	Latin America & the Caribbean	6.99	74.0	2.0	6.824288	5.679943	4.218635	6.748978	...	2.714233	
4	2020	Armenia	Caucasus & Central Asia	8.14	26.0	1.0	NaN	NaN	NaN	7.204175	...	5.170406	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3460	2000	Venezuela, RB	Latin America & the Caribbean	6.43	86.0	3.0	NaN	NaN	NaN	5.902894	...	6.417950	
3461	2000	Vietnam	South Asia	5.51	113.0	4.0	NaN	NaN	NaN	4.765274	...	NaN	
3462	2000	Yemen, Rep.	Middle East & North Africa	NaN	NaN	NaN	NaN	NaN	NaN	3.886318	...	NaN	
3463	2000	Zambia	Sub-Saharan Africa	7.03	72.0	3.0	NaN	NaN	NaN	6.087703	...	NaN	
3464	2000	Zimbabwe	Sub-Saharan Africa	5.35	116.0	4.0	NaN	NaN	NaN	5.150602	...	5.101020	

3465 rows × 141 columns

```
Index(['year', 'countries', 'region', 'hf_score', 'hf_rank', 'hf_quartile',
       'pf_rol_procedural', 'pf_rol_civil', 'pf_rol_criminal', 'pf_rol_vdem',
       ...
       'ef_regulation_business_adm', 'ef_regulation_business_burden',
       'ef_regulation_business_start', 'ef_regulation_business_impartial',
       'ef_regulation_business_licensing', 'ef_regulation_business_compliance',
       'ef_regulation_business', 'ef_regulation', 'ef_score', 'ef_rank'],
      dtype='object', length=141)
```

Fig 3.2.1 HFI Dataset

Fig 3.2.1 shows a snapshot of the Human Freedom index dataset and its features.

The Human Freedom Index (HFI) dataset, compiled by the Cato Institute, offers a comprehensive assessment of freedom levels across countries worldwide. It provides valuable insights into the state of personal, civil, and economic liberties, serving as a crucial resource for policymakers, researchers, and individuals concerned with understanding and promoting freedom.The HFI dataset adopts a multidimensional approach, employing a range of indicators to measure the overall freedom within each country. These indicators encompass key aspects such as the rule of law, property rights, freedom of expression and association, legal system integrity, sound money, trade openness, and regulatory environment.The HFI dataset holds immense significance for researchers, policymakers, social scientists, and advocates of liberty worldwide. Its comprehensive nature enables a detailed analysis of the relationships between freedom and various socio-economic indicators. Researchers can utilize the dataset to

investigate the impact of freedom on economic growth, human development, social well-being, and overall quality of life. Policymakers can draw insights from the dataset to inform policy decisions aimed at enhancing personal and economic freedoms within their respective nations. Furthermore, social scientists can leverage the HFI dataset to study the causes and consequences of variations in freedom across different countries and regions.The Human Freedom Index (HFI) dataset from Cato Institute uses a range of indicators and variables to measure the level of human freedom in various countries. The indicators fall under three main categories: Rule of Law, Security and Safety, and Movement.

The Rule of Law category includes variables such as the protection of property rights, the impartiality of the legal system, and the absence of corruption. The Security and Safety category measures the level of personal safety and security, including the use of torture, political imprisonment, and forced disappearance. The Movement category measures the level of freedom of movement, both within the country and across borders, including freedom of assembly and association, the right to emigrate and immigrate, and the right to work.

The dataset also includes several sub-variables to measure each of these indicators in more detail. For example, the Property Rights variable includes sub-variables such as the ability to own property, the ability to use property without interference, and the ability to transfer property. Similarly, the Personal Safety and Security variable includes sub-variables such as the level of violent crime, the level of terrorism, and the level of political violence.

Overall, the HFI dataset uses a comprehensive set of indicators and variables to measure the level of human freedom in different countries, allowing for detailed analysis of the factors that contribute to freedom and the areas where improvements are needed.

```
year                    float64
hf_score                float64
pf_rol                  float64
pf_ss                   float64
pf_movement             float64
pf_religion             float64
pf_association          float64
pf_expression           float64
pf_identity             float64
ef_government           float64
ef_legal                float64
ef_money                float64
ef_trade                float64
ef_trade_regulatory     float64
dtype: object
```

In [11]:
```
df2.isnull().sum()
```

Out[11]:
```
year                                  0
countries                             0
region                                0
hf_score                            382
hf_rank                             382
                                   ...
ef_regulation_business_compliance   422
ef_regulation_business              417
ef_regulation                       350
ef_score                            382
ef_rank                             382
Length: 141, dtype: int64
```

In [13]:
```
missing_df2 = (df2.isnull().sum() / len(df2)) * 100

print(missing_df2)
```

```
year                                 0.000000
countries                            0.000000
region                               0.000000
hf_score                            11.024531
hf_rank                             11.024531
                                       ...
ef_regulation_business_compliance   12.178932
ef_regulation_business              12.034632
ef_regulation                       10.101010
ef_score                            11.024531
ef_rank                             11.024531
Length: 141, dtype: float64
```

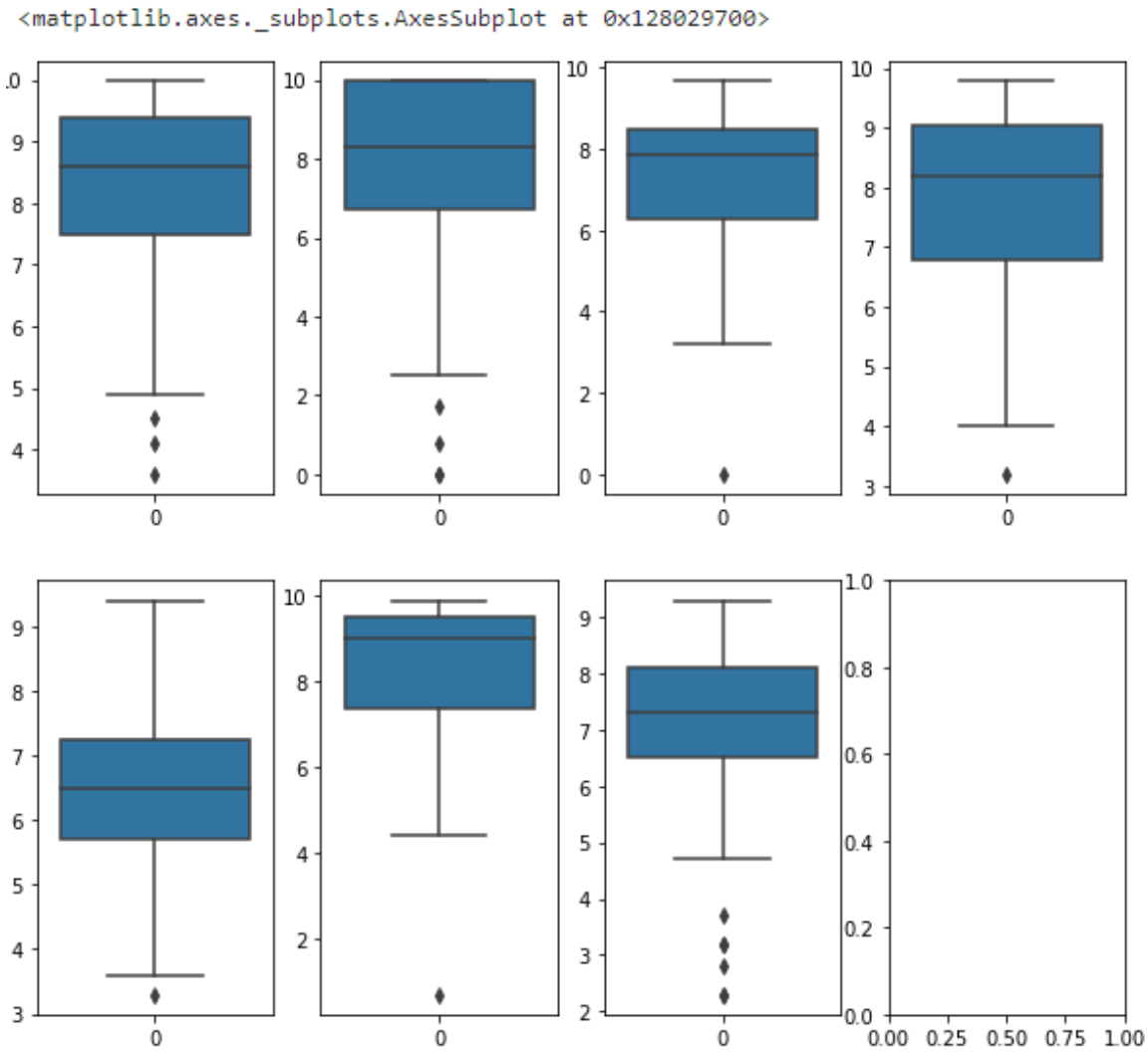


Fig 3.2.2

The data preprocessing phase involved various steps to ensure the quality and integrity of the dataset. Duplicated rows were identified and removed to eliminate redundancy. Missing values were addressed through mean imputation, where the average value of the respective feature was used to fill in the gaps. Features with a high percentage of missing values, exceeding 60%, were deemed unreliable and therefore removed from the dataset. Data types of specific columns were modified to ensure consistency and accuracy in subsequent analyses. Outliers, which could potentially skew the results, were detected and eliminated from the dataset. Additionally, measures were taken to remove noise from the data, enhancing the overall reliability and validity of the dataset for further analysis.

3.3 Data preprocessing for Global Terrorism Dataset

	eventid	iyear	imonth	iday	approxdate	extended	resolution	country	country_txt	region	...	addnotes	scite1	scite2	scite3	d
0	197000000001	1970	7	2	NaN	0	NaT	58	Dominican Republic	2	...	NaN	NaN	NaN	NaN	
1	197000000002	1970	0	0	NaN	0	NaT	130	Mexico	1	...	NaN	NaN	NaN	NaN	
2	197001000001	1970	1	0	NaN	0	NaT	160	Philippines	5	...	NaN	NaN	NaN	NaN	
3	197001000002	1970	1	0	NaN	0	NaT	78	Greece	8	...	NaN	NaN	NaN	NaN	
4	197001000003	1970	1	0	NaN	0	NaT	101	Japan	4	...	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
209701	202012310015	2020	12	31	2020-12-31 00:00:00	0	NaT	228	Yemen	10	...	NaN	"Al Houthi militia escalated in Hays and targe...	NaN	NaN	C
209702	202012310016	2020	12	31	2020-12-31 00:00:00	0	NaT	228	Yemen	10	...	NaN	"Al Houthi militia escalated in Hays and targe...	NaN	NaN	C
209703	202012310017	2020	12	31	NaN	0	NaT	75	Germany	8	...	NaN	"Far-left arson attack suspected on German asy...	"Fire of Bundeswehr vehicles in Leipzig, proba...	"Anarchist Antifa Take Credit for Arson Attack...	C
209704	202012310018	2020	12	31	NaN	0	NaT	4	Afghanistan	6	...	NaN	"Civil society activist and tribal elder kille...	"Terrorism Digest: 1-2 Jan 21," BBC Monitoring...	NaN	C
209705	202012310019	2020	12	31	NaN	1	NaT	33	Burkina Faso	11	...	NaN	"Terrorism Digest: 3-4 Jan 21," BBC Monitoring...	NaN	NaN	C

209706 rows × 135 columns

Fig 3.2.3 Global Terrorism Dataset

The Global Terrorism Dataset provides comprehensive information on terrorist attacks worldwide, offering valuable insights into the patterns, trends, and characteristics of terrorist activities. The dataset encompasses a wide range of variables, including the location, date, type of attack, target, and casualties involved in each incident. It covers a vast time span, allowing researchers to analyze the evolution of terrorism over the years. The dataset also includes information about the perpetrators, their affiliations, and the tactics employed. By examining this dataset, researchers can gain a deeper understanding of the dynamics of terrorism, identify hotspots, assess the effectiveness of counterterrorism measures, and contribute to the development of strategies aimed at preventing and mitigating terrorist threats.

The Global Terrorism Dataset encompasses a wide range of variables that provide comprehensive information on terrorist activities. These variables include the geographic location of each incident, allowing researchers to analyze regional patterns and identify areas prone to terrorism. The dataset also includes temporal variables such as the date and time of each attack, enabling the study of temporal trends and seasonal variations in terrorist activities. Furthermore, variables related to the type of attack, target, and weapons used

provide insights into the strategies and tactics employed by terrorist groups. Information on casualties, including the number of killed and injured, helps quantify

```
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)
```

```
In [14]: df1.duplicated()
Out[14]: 0         False
         1         False
         2         False
         3         False
         4         False
                   ...
         209701    False
         209702    False
         209703    False
         209704    False
         209705    False
         Length: 209706, dtype: bool

In [15]: df1.duplicated().sum()
Out[15]: 0
```

```
In [16]: df1.isnull().sum()
Out[16]: eventid            0
         iyear              0
         imonth             0
         iday               0
         approxdate    197017
                        ...
         INT_LOG            0
         INT_IDEO           0
         INT_MISC           0
         INT_ANY            0
         related       179102
         Length: 135, dtype: int64

In [17]: missing_df1 = (df1.isnull().sum() / len(df1)) * 100

         print(missing_df1)
         eventid        0.000000
         iyear          0.000000
         imonth         0.000000
         iday           0.000000
         approxdate    93.949148
                         ...
         INT_LOG        0.000000
         INT_IDEO       0.000000
         INT_MISC       0.000000
         INT_ANY        0.000000
         related       85.406235
         Length: 135, dtype: float64
```

Fig 2.3.4

In the preprocessing phase of the Global Terrorism Database, several data cleaning techniques were applied. Duplicated rows were identified and

removed to ensure the dataset's integrity and avoid duplication of information. Missing values were addressed by employing mean imputation, where the average value of the available data was used to fill in the gaps. Additionally, features with a substantial number of missing values, exceeding the threshold of 60%, were removed from the dataset to maintain data quality. Data types of specific columns were modified to ensure consistency and compatibility for further analysis.

3.4 Data exploration and visualization

In this project, data ingestion was performed using Amazon S3. Two buckets were created and the datasets were ingested into them. The datasets were then transformed and joined using AWS Glue. The datasets were merged based on the 'country_text' column, which was a common column in the datasets. The resultant merged dataset was saved in another bucket. This process ensured that the datasets were properly prepared for the subsequent data processing, analysis, and visualization stages.

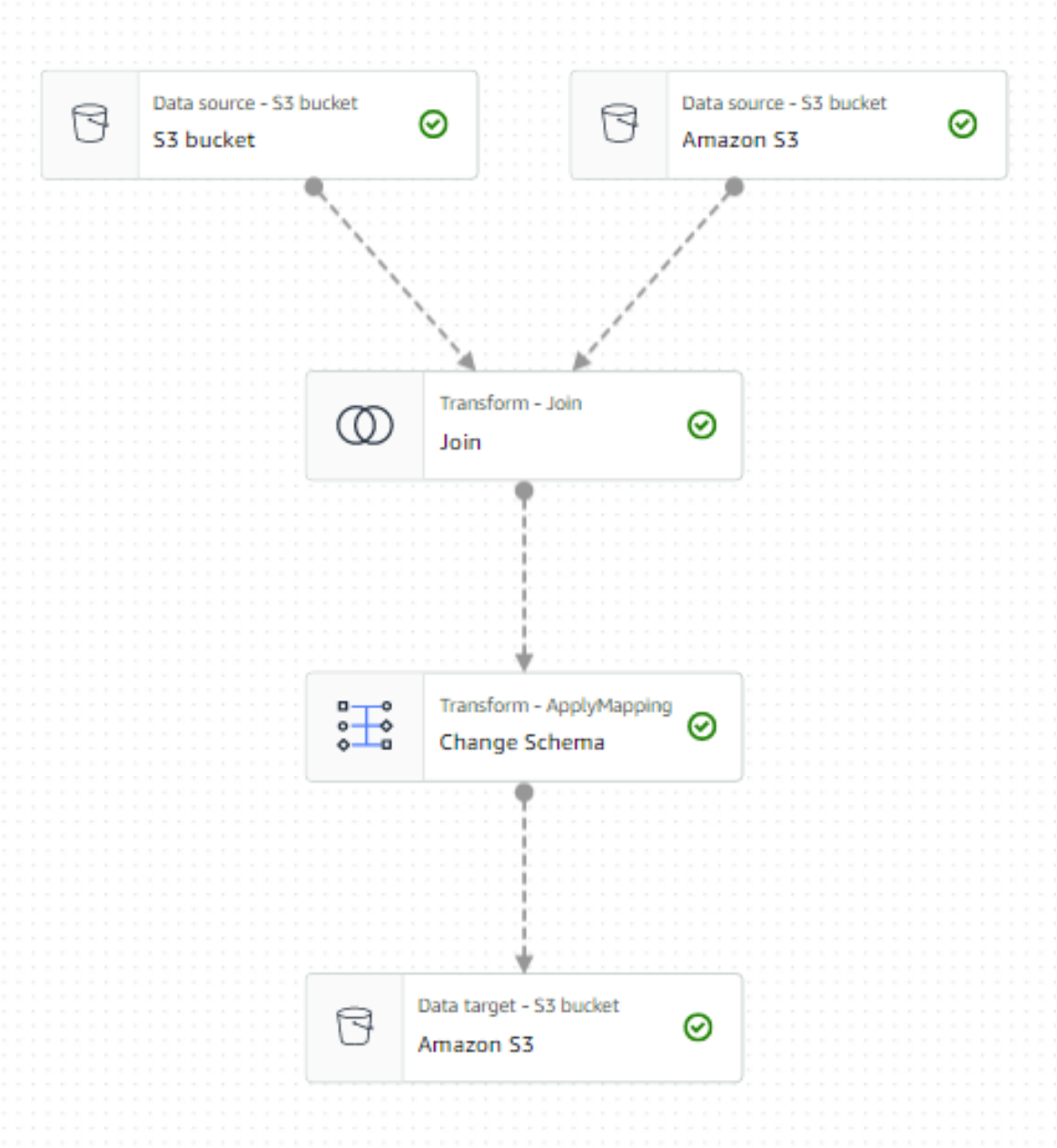

Data source - S3 bucket
S3 bucket
Data source - S3 bucket
Amazon S3
Transform - Join
Join
Transform - ApplyMapping
Change Schema
Data target - S3 bucket
Amazon S3

Fig 3.4.1

	eventid	iyear	imonth	iday	approxdate	extended	resolution	country	country_txt	region_x	...	ef_regulation_business_adm	ef_regulation_bu
0	197000000001	1970	7	2	NaN	0	NaT	58	Dominican Republic	2	...	3.270430	
1	197000000001	1970	7	2	NaN	0	NaT	58	Dominican Republic	2	...	3.270430	
2	197000000001	1970	7	2	NaN	0	NaT	58	Dominican Republic	2	...	3.270430	
3	197000000001	1970	7	2	NaN	0	NaT	58	Dominican Republic	2	...	3.115157	
4	197000000001	1970	7	2	NaN	0	NaT	58	Dominican Republic	2	...	2.805399	
...	...	...	...	...	...	...	...	...	...	...	...	...	
3554203	201812120020	2018	12	12	NaN	0	NaT	1001	Serbia	9	...	NaN	
3554204	201812120020	2018	12	12	NaN	0	NaT	1001	Serbia	9	...	NaN	
3554205	201812120020	2018	12	12	NaN	0	NaT	1001	Serbia	9	...	NaN	
3554206	201812120020	2018	12	12	NaN	0	NaT	1001	Serbia	9	...	NaN	
3554207	201812120020	2018	12	12	NaN	0	NaT	1001	Serbia	9	...	NaN	

3554208 rows × 275 columns

Fig 3.4.2

Figure 3.4.2 illustrates the merged dataset obtained from the data ingestion and transformation process. Upon obtaining the merged dataset, exploratory data analysis (EDA) was conducted to gain insights into the data. EDA techniques such as descriptive statistics, data visualization, and pattern identification were employed to understand the characteristics and relationships within the dataset. This step allowed for a comprehensive exploration of the data, enabling the identification of key trends, patterns, and potential areas of interest for further analysis.

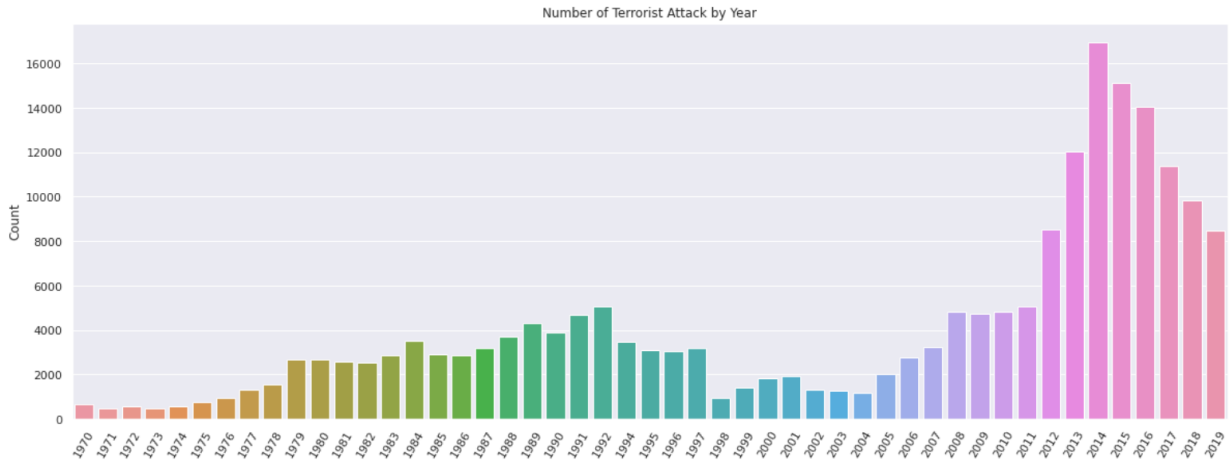


Fig 3.4.3

Figure 3.4.3 presents the number of terrorist attacks by year, providing a visual representation of the data. The figure clearly demonstrates a notable increase in the frequency of terrorist attacks following the year 2011, which coincides with the significant event of the 9/11 attacks in the United States. This observation suggests a potential correlation between the occurrence of

the 9/11 attacks and the subsequent rise in global terrorism incidents. The visual depiction offers a concise and impactful representation of this temporal trend, facilitating a clear understanding of the dataset's dynamics and highlighting the significance of the post-2011 period in terms of terrorist activity.

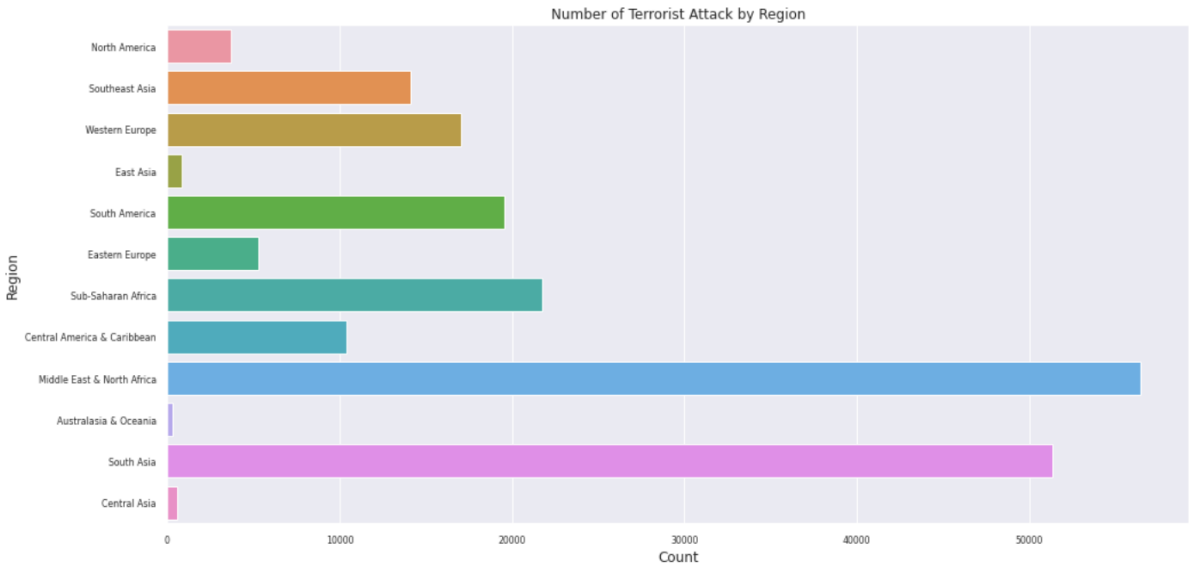


Fig 3.4.4

Figure 3.4.4 displays the number of attacks categorized by region. The data reveals that the regions of Middle East and North Africa, along with South Asia, exhibit the highest frequency of attacks. Conversely, the regions of Oceania and Central Asia display the lowest number of attacks. This graphical representation provides a comprehensive overview of the distribution of attacks across different regions, highlighting the contrasting levels of terrorist activity in various parts of the world. The figure serves as a valuable visual aid in identifying the regions with the highest and lowest incidences of attacks, contributing to a deeper understanding of the global patterns and regional variations in terrorism.

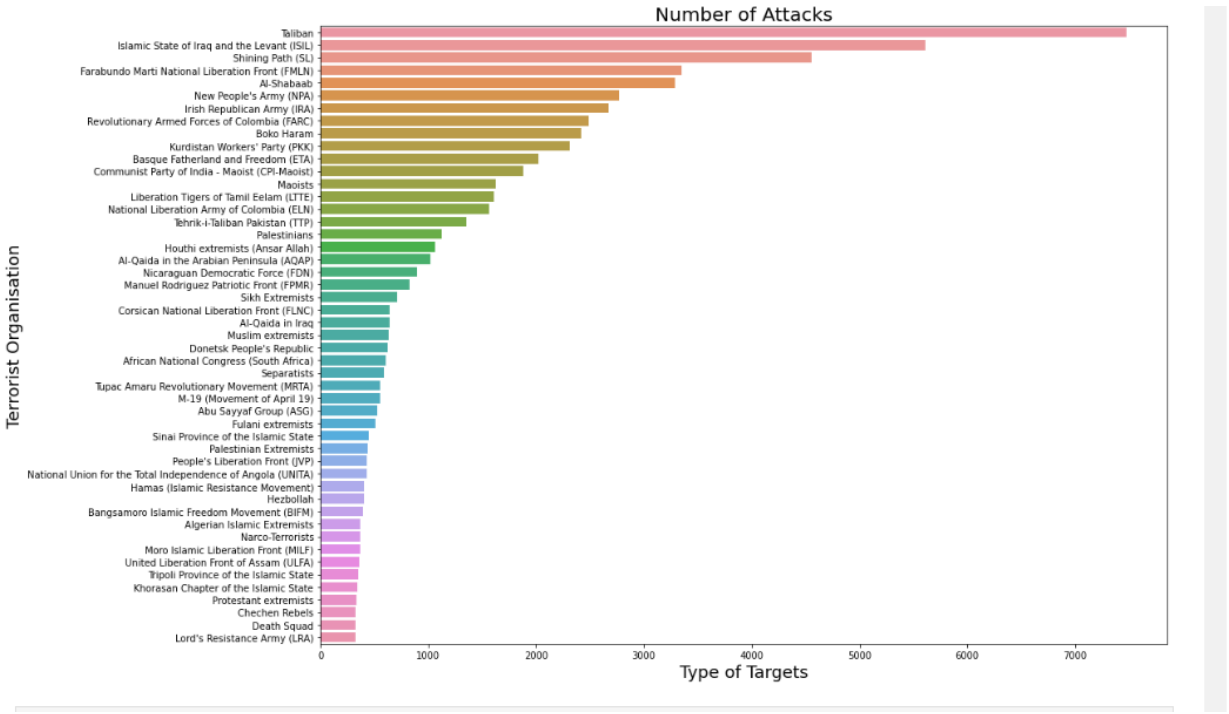


Fig 3.4.5

Figure 3.4.5 presents the number of attacks attributed to various terrorist organizations. The data clearly indicates that the Taliban organization has the highest count of attacks, signifying their significant involvement in terrorist activities. On the other hand, the terrorist groups known as Death Quad and Lord's Resistance Army (LRA) have the lowest number of recorded attacks. This visual representation allows for a quick comparison of the frequency of attacks carried out by different terrorist organizations, emphasizing the dominance of the Taliban and the relatively limited engagement of the Death Quad and LRA. The figure plays a crucial role in identifying the major perpetrators of terrorist acts and their relative impact in terms of attack incidents.

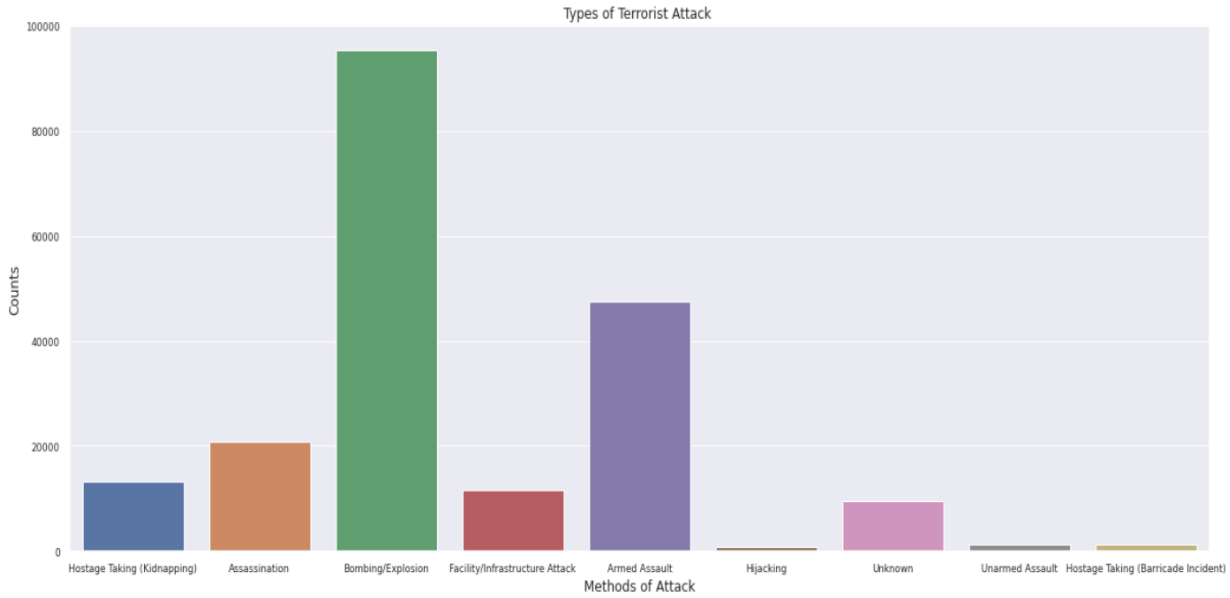


Fig 3.4.6

Figure 3.4.6 illustrates the distribution of different types of attacks carried out by terrorist organizations. The data reveals that bombing or explosion incidents were the most prevalent type of attack perpetrated by these groups. This indicates their propensity for utilizing explosives as a means of causing destruction and spreading fear. On the other hand, hijacking and unarmed assault were the least frequently observed attack types. This visual representation offers insights into the preferred methods employed by terrorist organizations, highlighting their focus on bombings while demonstrating a lower occurrence of hijackings and unarmed assaults. Understanding the prevalent attack types is crucial for developing counterterrorism strategies and mitigating the risks associated with specific attack methodologies.

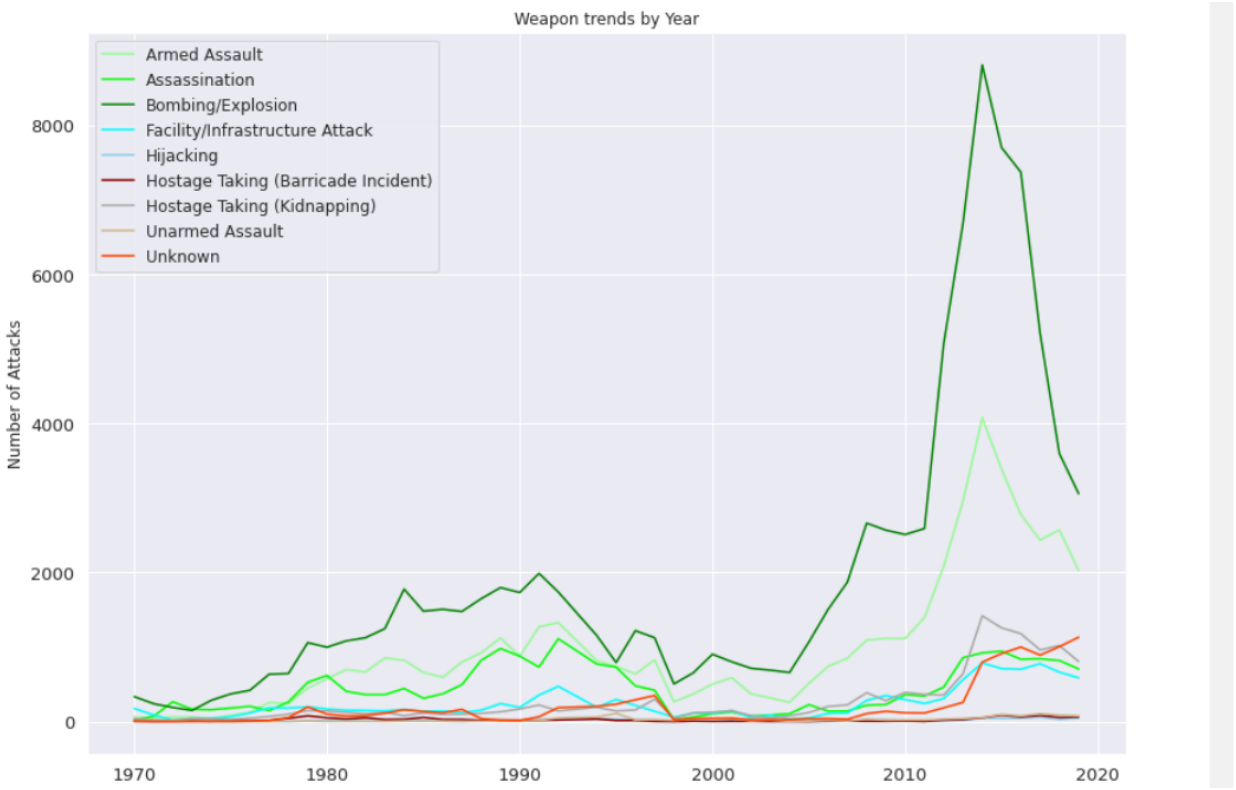


Fig 3.4.7

Figure 3.4.7 depicts a line plot illustrating the frequency of different types of attacks over the years. The data reveals that after the year 2010, there was a significant increase in the occurrences of bombing or explosion incidents, which became the most prevalent type of attack. This notable rise coincides with the historical event of 9/11, suggesting a potential correlation between the increased prevalence of such attacks and the impact of that event. The line plot provides a visual representation of the upward trend in bombing or explosion incidents, emphasizing the shift in attack patterns post-2010. This insight underscores the need for enhanced counterterrorism measures to address the specific challenges posed by this particular type of attack.

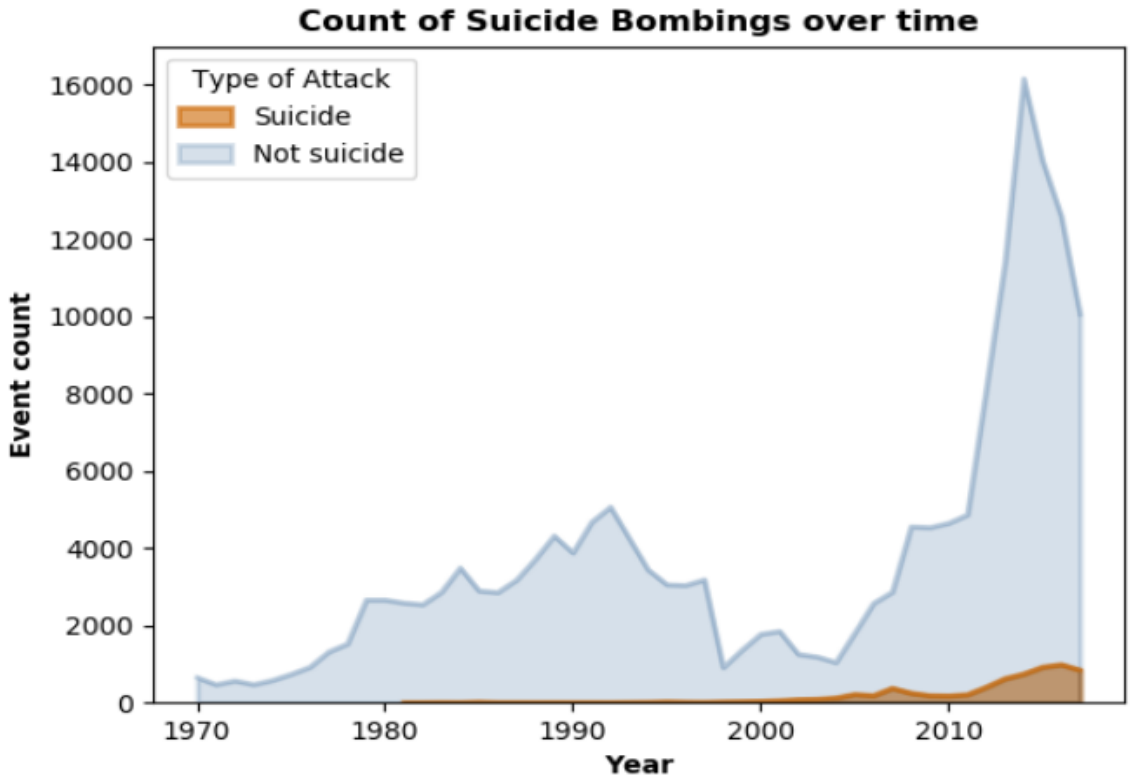


Fig 3.4.8

Figure 3.4.8 presents a density plot that provides further evidence of the increase in suicide bombings after the year 2010, coinciding with the significant event of 9/11. The density plot showcases the distribution of suicide bombing incidents over time, emphasizing the shift in frequency towards higher values post-2010. The plot's shape reveals a notable concentration of incidents in the later years, indicating an upward trend in the occurrence of suicide bombings. This finding strengthens the observation made in previous analyses, highlighting the correlation between the post-2010 period and the heightened prevalence of this specific type of attack. The density plot serves as a visual confirmation, reinforcing the need for focused efforts and counterterrorism strategies to address the escalating threat of suicide bombings.

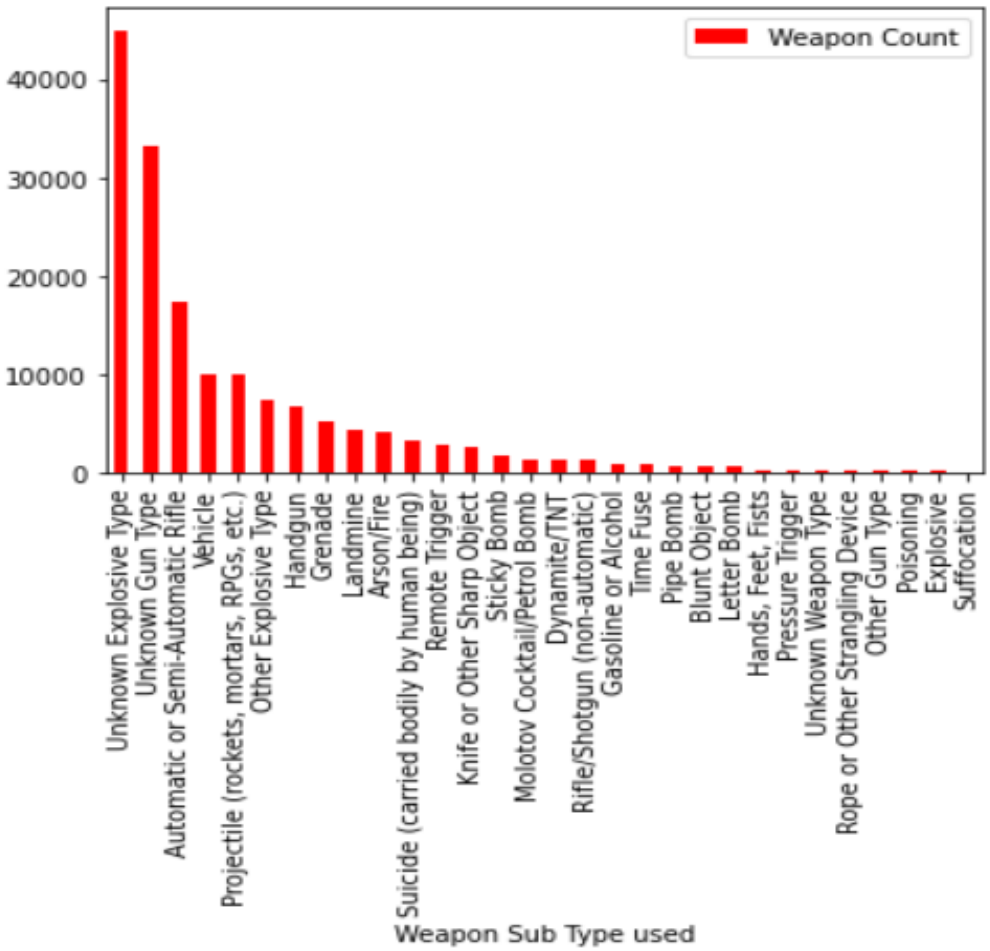


Fig 3.4.9

Figure 3.4.9 displays a bar chart representing the distribution of weapon types used in terrorist attacks. The chart highlights that the category "unknown explosive" is the most commonly used weapon, indicating that the specific nature of the explosive device was unidentified or undisclosed in the dataset. This suggests a significant portion of attacks involved explosives whose exact composition or origin was not documented. On the other hand, the bar chart also reveals that suffocation and poisoning were among the least prevalent weapon types employed in terrorist incidents. These findings provide insights into the diverse range of weapons utilized by terrorist organizations, with explosives being the most frequently deployed, while suffocation and poisoning are relatively uncommon methods. Understanding the prevalent weapon types aids in assessing the tactics employed by terrorist groups and enables policymakers and security agencies to tailor effective counterterrorism strategies accordingly.

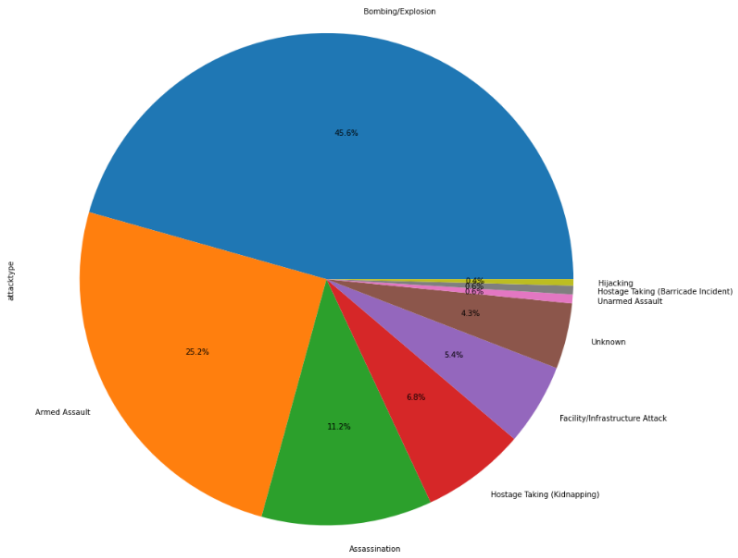


Fig 3.4.10

Figure 3.4.10 presents a pie chart depicting the distribution of terrorist attack types. The chart reaffirms our previous analysis, highlighting that bombing constitutes the largest share of total attacks. This indicates that bombings, including explosive devices and other forms of detonations, are the most prevalent method employed by terrorist organizations. The pie chart visually emphasizes the significance of bombing as a preferred tactic, as it accounts for a substantial portion of the overall attack landscape. The presence of other attack types, such as armed assault, assassination, and hostage-taking, is comparatively smaller in proportion. This reinforces the understanding that bombings play a central role in terrorist activities globally. Such insights assist in comprehending the dominant strategies utilized by terrorist groups and inform efforts to develop effective counterterrorism measures focused on mitigating the risks associated with bombings.

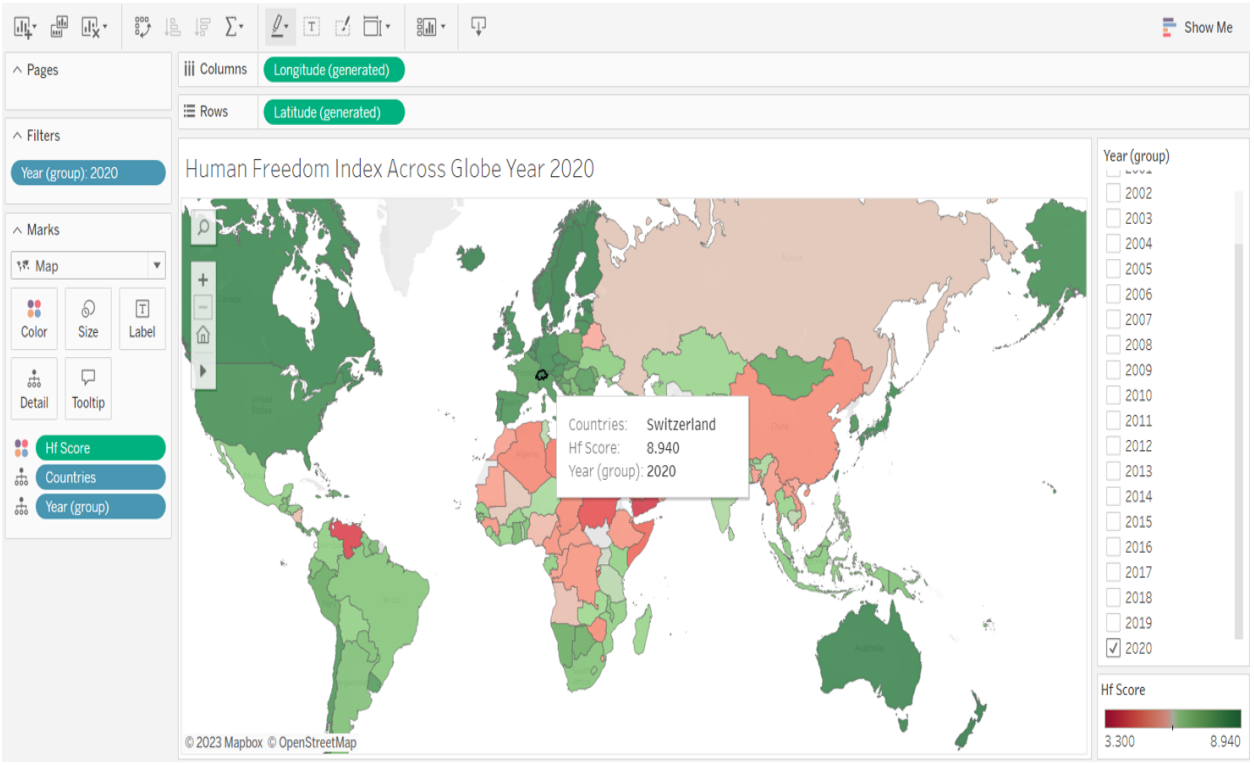


Fig 3.4.11

Figure 3.4.11 showcases a choropleth map representing the Human Freedom Index (HFI) scores across the globe. The map provides a visual representation of the variations in HFI scores among different countries. It is evident from the map that Switzerland has the highest HFI score, indicating a greater degree of human freedom in terms of personal, civil, and economic liberties. On the other hand, Syria exhibits the lowest HFI score, indicating a significant lack of freedom in various aspects of life. This aligns with our previous analysis, which indicated that the Middle East region, including countries like Syria, experiences a higher frequency of terrorist attacks. The correlation between lower HFI scores and increased terrorist activity in the Middle East reinforces the notion that socio-political conditions and restrictions on freedoms may contribute to the prevalence of terrorism in certain regions. The choropleth map provides valuable insights into the distribution of HFI scores worldwide and their potential relationship with terrorist incidents.

4 Machine Learning to Predict Successful Attack

In order to predict the success of an attack using machine learning techniques, a new target class called "success" is defined. This target class is determined based on the number of people killed in an attack. If the number of people killed is greater than 0, the output value is set to 1, indicating a successful attack. Conversely, if the number of people killed is 0 or less, the output

value is set to 0, indicating an unsuccessful attack. This binary classification approach allows for the development of a predictive model that can identify the likelihood of a successful attack based on the given features and the outcome of casualties.

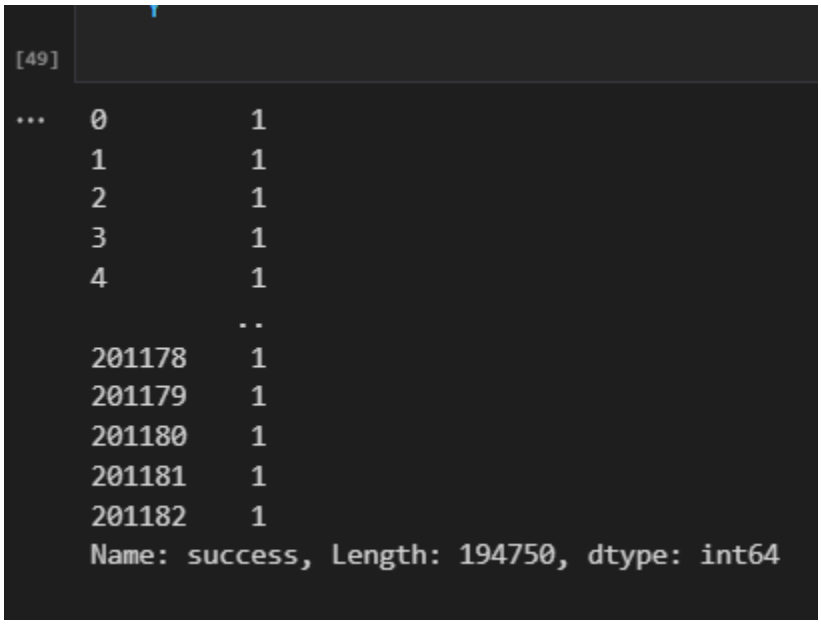


Fig 4.1Target Class

4.1 Feature Selection and Extraction

Principal Component Analysis (PCA) is a dimensionality reduction technique commonly used in machine learning and data analysis. Its primary goal is to transform a high-dimensional dataset into a lower-dimensional space while retaining most of the important information. PCA achieves this by identifying the directions (principal components) along which the data varies the most.

One of the key applications of PCA is feature selection. By performing PCA on a dataset, we can obtain a new set of orthogonal features called principal components. These components are ranked based on their contribution to the variance in the data. The higher the variance explained by a principal component, the more important it is considered.

Random Forest, on the other hand, is an ensemble learning algorithm that combines multiple decision trees to make predictions. It has the ability to measure the importance of features in the dataset based on their impact on the accuracy of the model. This information can be utilized for feature selection.

To perform feature selection using Random Forest, we can calculate the feature importances provided by the algorithm. These importances indicate the relative importance of each feature in determining the outcome. Features with higher importances are considered more influential in the model's predictions.

By combining PCA and feature selection with Random Forest, we can select a subset of the most informative features from the dataset. This approach helps to reduce dimensionality, eliminate irrelevant or redundant features, and improve the model's predictive performance. By retaining only the most important features, we can simplify the model, reduce overfitting, and enhance interpretability

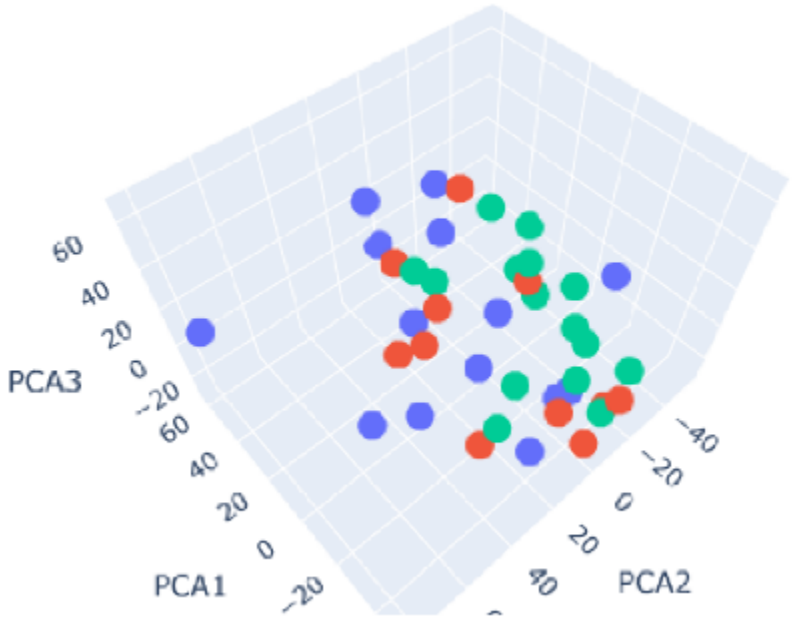


Fig 4.2

PCA1 has variance ratio of 0.4 indicates that PCA1 carries substantial information from our dataset. It suggests that this component alone can capture a considerable amount of the data's patterns, trends, or variations.

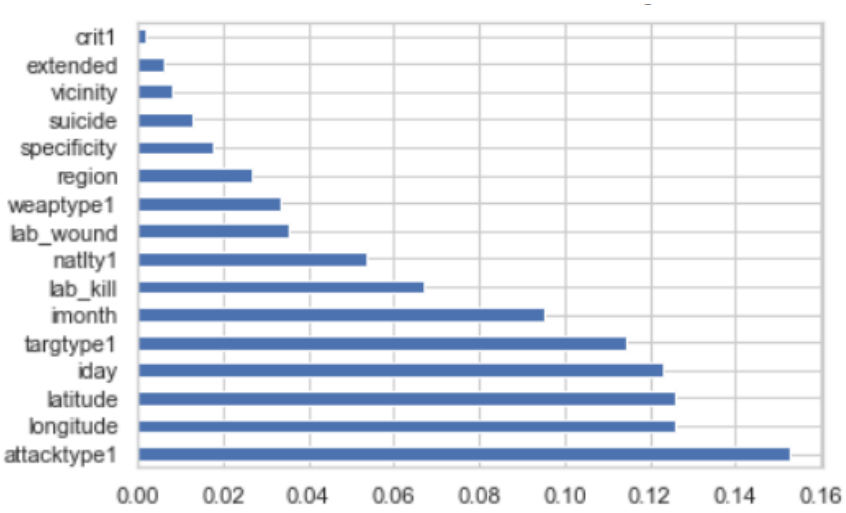


Fig 4.3

Two machine learning algorithms, namely Random Forest and Decision Tree, were employed for the prediction of successful attacks. These algorithms are well-known and widely used in the field of machine learning for classification tasks. The Random Forest algorithm is an ensemble learning method that constructs multiple decision trees and combines their predictions to make a final prediction. On the other hand, the Decision Tree algorithm creates a tree-like model that represents decisions and their possible consequences. By utilizing these algorithms, the aim is to leverage their ability to learn patterns and relationships from the provided dataset in order to accurately classify attacks as successful or unsuccessful based on the selected features.

4.2 Models

4.2.1 Decision Tree

A Decision Tree is a popular supervised machine learning algorithm that is widely used for classification and regression tasks. It creates a hierarchical tree-like structure where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label or a prediction.

The decision-making process in a Decision Tree starts at the root node and follows a path down the tree based on the feature tests. The splitting of nodes is determined by selecting the feature that best separates the data based on certain criteria such as Gini impurity or information gain. This process is repeated recursively until a stopping criterion is met, such as reaching a maximum depth or having homogeneous class labels in the leaf nodes.

Decision Trees offer several advantages, including their interpretability and ease of understanding. They provide explicit rules that can be easily

visualized and explained, making them useful in decision-making processes. Decision Trees can handle both numerical and categorical data and can automatically handle missing values and outliers. They are also robust to irrelevant features, as they only consider the relevant ones for making decisions.

However, Decision Trees can be prone to overfitting, especially when the tree becomes too deep or complex. To mitigate this issue, techniques like pruning, limiting the maximum depth, or setting a minimum number of samples for splitting can be employed. Additionally, Decision Trees may suffer from high variance and instability, meaning that small changes in the data can lead to different tree structures and predictions.

4.2.2 Random forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. It is a powerful and widely used algorithm for classification and regression tasks. The key idea behind Random Forest is to introduce randomness into the decision tree construction process to reduce overfitting and improve generalization.

Random Forest works by creating an ensemble of decision trees, where each tree is trained on a random subset of the original dataset (bootstrap sampling) and only considers a random subset of features at each split. These randomization techniques ensure diversity among the trees and reduce the correlation between them.

When making predictions, Random Forest aggregates the predictions of all individual trees and outputs the majority class (for classification) or the average (for regression). This ensemble approach provides more robust and accurate predictions compared to a single decision tree.

Random Forest offers several advantages. It can handle large datasets with high dimensionality and is less prone to overfitting compared to a single decision tree. It can capture complex relationships between features and target variables and provide estimates of feature importance, allowing for feature selection. Random Forest is also computationally efficient, as the training process can be parallelized.

However, Random Forest has some limitations. It can be challenging to interpret and explain compared to a single decision tree. It may also require careful tuning of hyperparameters, such as the number of trees in the forest and the maximum depth of each tree. Additionally, Random Forest may not perform well on datasets with imbalanced class distributions or when dealing with noisy data.

In summary, Random Forest is a versatile ensemble learning algorithm that leverages the power of decision trees for robust and accurate predictions. Its ability to reduce overfitting, handle high-dimensional data, and provide feature importance makes it a popular choice in various machine learning applications.

4.2.3 Performance

The classification report is a comprehensive performance metric used to evaluate the performance of a classification model. It provides a detailed breakdown of various evaluation metrics such as precision, recall, F1-score, and support for each class in the dataset. Precision represents the proportion of correctly predicted positive instances out of all instances predicted as positive, measuring the model's accuracy in identifying true positives. Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances out of all actual positive instances, indicating the model's ability to identify all relevant positive instances. The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both precision and recall. It is useful when the dataset is imbalanced or when there is a trade-off between precision and recall. Additionally, the support metric represents the number of instances in each class. The classification report offers a concise summary of the model's performance for each class, enabling a comprehensive evaluation of the model's effectiveness in classifying different classes and identifying potential biases or weaknesses in the model's predictions.The classification report is a valuable tool in assessing the performance of a classification model. It provides a detailed analysis of various performance metrics that help in understanding the model's behavior and effectiveness in predicting different classes. One of the key metrics in the classification report is precision, which measures the proportion of correctly predicted positive instances out of all instances predicted as positive. A high precision score indicates that the model has a low false positive rate, meaning it is accurately identifying positive instances.

Another important metric in the classification report is recall, also known as sensitivity or true positive rate. Recall measures the proportion of correctly predicted positive instances out of all actual positive instances. It assesses the model's ability to capture all relevant positive instances and avoid false negatives. A high recall score indicates that the model is effectively identifying positive instances and minimizing the false negative rate.

The F1-score is a combined measure that takes into account both precision and recall. It is the harmonic mean of the two metrics and provides a balanced assessment of the model's performance. The F1-score is particularly useful when there is a trade-off between precision and recall, or when the dataset is imbalanced, as it considers both the model's ability to minimize false positives and false negatives.

Additionally, the classification report includes the support metric, which represents the number of instances in each class. This metric provides insights into the distribution of the classes in the dataset and helps in identifying potential biases or imbalances.

By examining the classification report, one can gain a comprehensive understanding of the model's strengths and weaknesses in classifying different classes. It allows for an in-depth evaluation of the model's performance, enabling researchers and practitioners to make informed decisions regarding the suitability and reliability of the classification model for their specific task or problem domain.

	precision	recall	f1-score	support
0	0.91	1.00	0.95	71
1	1.00	0.84	0.91	43
accuracy			0.94	114
macro avg	0.96	0.92	0.93	114
weighted avg	0.94	0.94	0.94	114

Fig 4.2.3. Decision Tree

	precision	recall	f1-score	support
0	0.96	1.00	0.98	71
1	1.00	0.93	0.96	43
accuracy			0.97	114
macro avg	0.98	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Fig 4.2.3.1 Random Forest

For the Decision Tree model, the accuracy is 0.94, indicating that the model correctly predicts the target class in 94% of the cases. The weighted average F1-score, which considers both precision and recall, is also 0.94. This indicates that the model achieves a good balance between precision and recall for the overall classification task.

When specifically looking at the target class 0, the F1-score is 0.91. This means that the model performs well in predicting instances belonging to class 0, with a harmonious balance between precision and recall. For the target class 1, the F1-score is 0.94, indicating that the model effectively identifies instances belonging to class 1 with a high level of accuracy.

Moving on to the Random Forest model, the accuracy is higher at 0.97, demonstrating an improved performance compared to the Decision Tree model. The weighted average F1-score

is also higher at 0.97, suggesting that the Random Forest model achieves a more balanced and accurate classification across all classes.

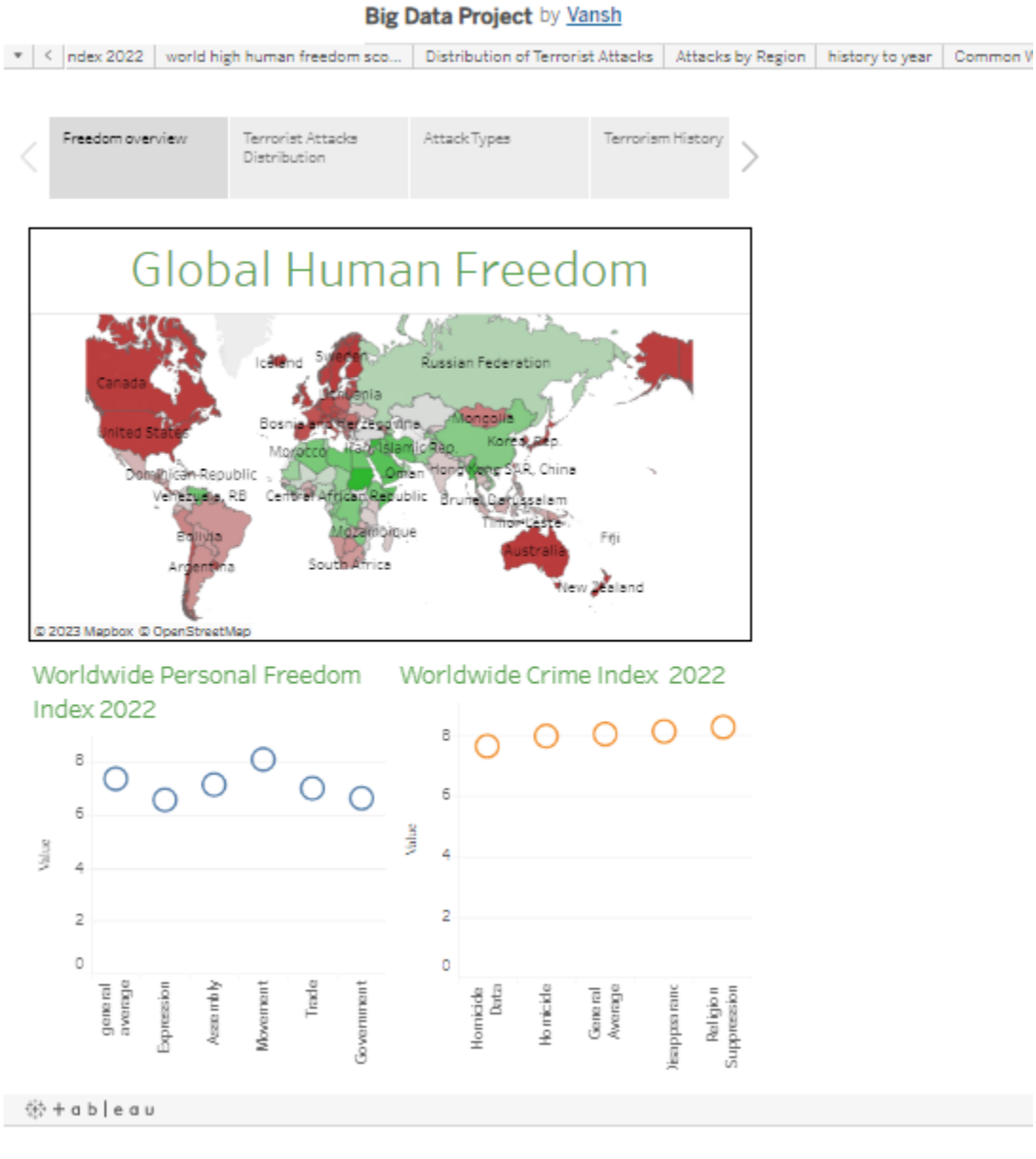
For class 0, the F1-score is 0.98, indicating an excellent performance in predicting instances belonging to this class. Similarly, for class 1, the F1-score is 0.96, suggesting a strong ability to identify instances in this class accurately.

Overall, both the Decision Tree and Random Forest models exhibit strong performance, with the Random Forest model achieving higher accuracy and F1-scores. These results indicate that the models are effective in classifying the dataset and can provide reliable predictions for the target variable.

The Random Forest model achieves a slightly higher accuracy of 0.97 compared to the Decision Tree model's accuracy of 0.94. Similarly, the weighted average F1-score for the Random Forest model is 0.97, indicating a more balanced and accurate classification across all classes, while the Decision Tree model has a weighted average F1-score of 0.94.Moreover, when looking at the F1-scores for individual classes, the Random Forest model shows higher scores with an F1-score of 0.98 for class 0 and 0.96 for class 1, compared to the Decision Tree model's F1-score of 0.91 for class 0 and 0.94 for class 1.

These results suggest that the Random Forest model outperforms the Decision Tree model in terms of accuracy and overall classification performance. It demonstrates higher accuracy, better balance between precision and recall, and more accurate predictions for both classes. Therefore, based on the provided metrics, the Random Forest model is considered to be better in this scenario.

5 Interactive Dashboard



Dashboard::https://public.tableau.com/views/BigDataProject_16838508964770/TerrorismAnalysis?:language=en-US&:display_count=n&:origin=viz_share_link

6 Conclusion

- Distribution of terroristic attacks seem to be focus on major cities more than country side or rural areas

- Looking at the history and how terroristic behavior change over time, it seems to be influenced by social and historical events rather than a general trait or behavior in people

- The number of wounded are always higher than the number of killed, with an apparent anomaly by recent year. The anomaly could be due to changing of counting ways and methods of gathering data, or due to the increase number of events (a point for further investigation)

- The primary weapon of attack is explosives by different kinds, then guns of different kinds, secondary ones are projectiles and blunt weapons and nearly nonexistent is strangling of different kinds

- No clear definition is given to separate terrorism from general crimes; that's why other studies and data gathered on crimes may have different results even if it addresses the same points

- From the freedom dataset, it seems that the general world's average of "freedom" is about 8, which is good, but with a room of improvement

- In individual freedoms, freedom of movement seems to be the least restricted of all, sometimes having a score even high than average

- Finally, freedom of definition and biases in it seems to play a big role on the appearance of data, favoring countries that set the definition and gather the data; which would be the same if it was conducted by other countries, favoring them over the others..... Politics and political influence seems to play a big role in international research and have influence over it, which hinders objective oriented research.Which is a very important metric of freedom, that requires further studying

References

Choi, W., & Choi, W. S. (2008). Economic sanctions, poverty, and international terrorism. Journal of Conflict Resolution, 52(1), 96-116. doi:10.1177/0022002707310447

Krueger, A. B., & Malečková, J. (2003). Terrorism and poverty: Evidence from the Middle East. Journal of Economic Perspectives, 17(4), 119-144. doi:10.1257/089533003772034925

Smith, J., Nair, R., & Salerno, J. (2019). Machine learning approaches for counterterrorism. Security Informatics, 8(1), 3. doi:10.1186/s13388-019-0019-0

Bandyopadhyay, S., & Younas, J. (2011). Terrorism and foreign direct investment: An empirical analysis of regional differences in India. Review of Development Economics, 15(3), 504-516. doi:10.1111/j.1467-9361.2011.00607.x