CS 657 Mining Massive Datasets
Final Project: Global Terrorism Predictive Analytics

Ritwik Sharma (G01373222) and Mukund Sharma (G01374620)
Graduate Student, Computer Science
George Mason University
Fairfax, Virginia, United States

ABSTRACT
The purpose of this project report is to employ geospatial, network and survival analyses, anomaly detection, predictive modeling, topic modeling, and political stability analysis to comprehensively analyze terrorism incidents, offering a data-driven and user-friendly approach for understanding patterns, relationships, durations, anomalies, predictions, themes, and political influences.

- **Keywords:** PySpark, Geospatial Analysis, Network Analysis, Survival Analysis, Anomaly Detection, Predictive Modeling, Topic Modeling, Political Stability

1. INTRODUCTION

I. Global Terrorism Database (GTD)

The Global Terrorism Database (GTD) is a comprehensive database of terrorist incidents from 1970 to 2021, maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland, College Park.

II. World Bank Data and Political Stability

The World Bank Data refers to a vast collection of global development indicators, financial and economic data, compiled and maintained by the World Bank Group, offering a comprehensive resource for analyzing and understanding global trends, policies, and socioeconomic dynamics across countries and regions.

The Political Stability dataset from World Data Bank provides aggregate and individual governance indicators for six dimensions of governance, including Voice and Accountability, Political Stability and Absence of Violence/Terrorism, Government Effectiveness, Regulatory Quality, Rule of Law, and Control of Corruption from 1996 to 2022. Political stability index ranges from -2.5 (indicating weakness) to 2.5 (representing strength). In 2021, the global average across 193 countries stood at -0.07 points for the index.

III. Project Methodologies

The project encapsulates a comprehensive exploration and analysis of terrorism incidents employing a diverse range of analytical methodologies:

- **Geospatial Analysis**: Perform geospatial analysis to map and analyze terrorism incidents over time, identifying patterns and high-risk areas.
- **Network Analysis**: Conduct network analysis to map relationships among terrorist groups, locations, and targets, uncovering key clusters, connections, and central nodes to gain insights into group operations and recruitment.
- **Survival Analysis**: Utilize survival analysis to model the duration of terror campaigns and group

lifespans, revealing patterns in terrorism events over time.

- **Anomaly Detection**: Implement anomaly detection techniques to identify unusual patterns or outliers in terrorism data, potentially signifying emerging threats or significant events necessitating intervention.
- **Topic Modeling**: Employ topic modeling to discover prevalent themes and subjects within terrorist communications, focusing on attack summaries, motive and ransom note to gain deeper insights into their tactics.
- **Predictive Modeling**: Develop a predictive model to forecast the likelihood of future terrorist incidents across regions, countries, states, and cities by analyzing attack, target, and weapon types, as well as the nationality of terrorists and organizations, facilitating targeted allocation of counter-terrorism resources.
- **Political Stability Analysis**: Investigate the relationship between terrorism activity and the political stability of countries to assess how political conditions may influence terrorist incidents.

2. LITERATURE SURVEY

I. Apache Spark

It is an open-source distributed general-purpose cluster-computing framework that is designed for efficient processing of large datasets. Spark utilizes in-memory caching and optimized query execution to provide superior performance compared to traditional disk-based big data frameworks.

PySpark

It is the Python API for Spark that allows programmers to harness the power of Spark from Python applications and shells. PySpark supports much of the same operations as standard Spark, including distributed datasets, lambda functions, and lazy evaluation.

II. Topic Modeling and Text Classification

- **Topic modeling:** It is a natural language processing technique that aims to identify the underlying topics in a collection of text documents. It has been widely used in various text classification tasks to uncover the latent themes or topics within the text data.
- **LDA (Latent Dirichlet Allocation):** It is a probabilistic model that uncovers topics in a collection of documents.
- **Log Perplexity:** It is a metric used for evaluating the quality of topic modeling in Latent Dirichlet Allocation (LDA) models, calculated as the negative log-likelihood of a held-out test dataset normalized by the total number of words.

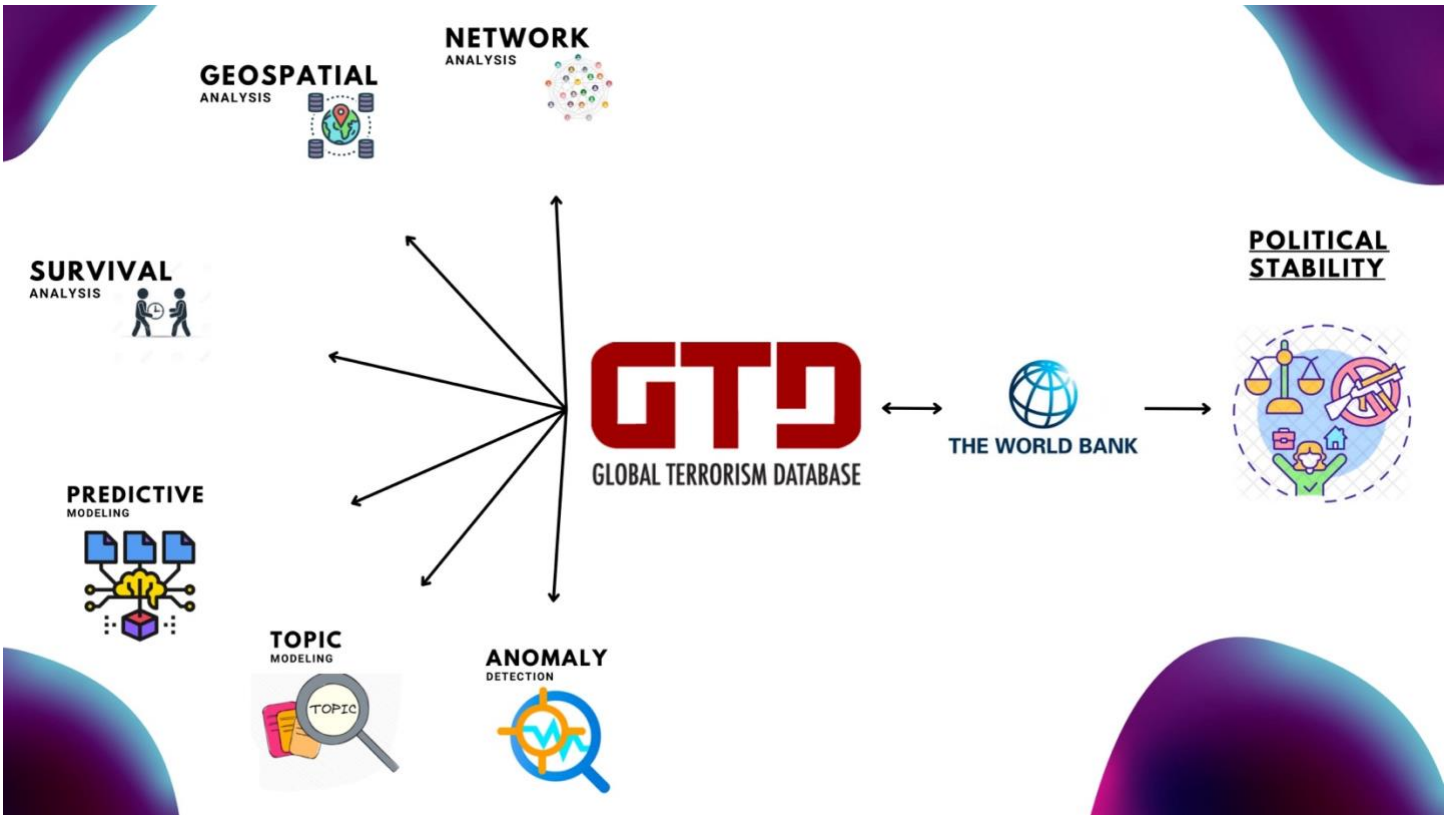
III. XGBoost

XGBoost is an ensemble learning method that utilizes gradient boosting to sequentially combine weak learners, creating a powerful and efficient boosting algorithm known for its speed and predictive accuracy. Its algorithm focuses on minimizing the overall errors by optimizing a predefined loss function through gradient descent and tree-based models, enhancing performance through boosting iterations.

IV. Cross-Validation, Hyperparameter Tuning and Evaluation Metrics

- **Cross-Validation:** It is a technique that assesses a model's generalization performance by partitioning the dataset into multiple subsets, training on a portion, and validating on the rest, ensuring robustness against overfitting.
- **Hyperparameter Tuning:** It involves optimizing the configuration settings of a machine learning model to achieve best possible performance, often through techniques like grid search or randomized search.
- **Evaluation Metrics:** It provides quantitative measures for assessing a model's performance in classification tasks.
- **ROC:** The Receiver Operating Characteristic (ROC) score is a graphical representation of a model's ability to distinguish between classes in a binary classification problem.

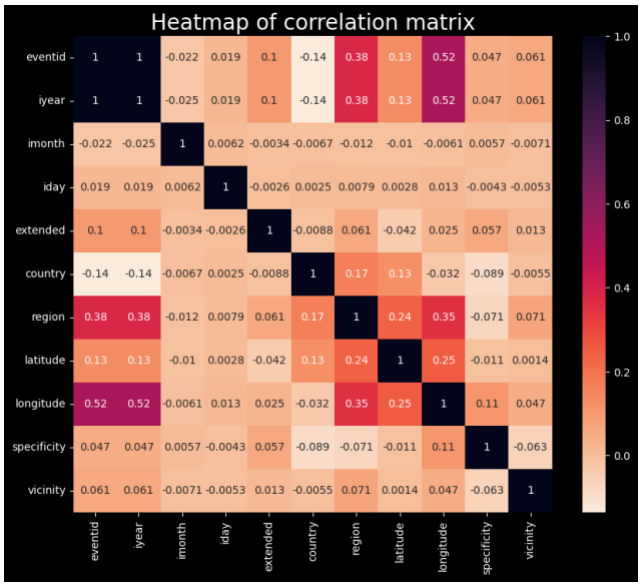
3. ARCHITECTURE

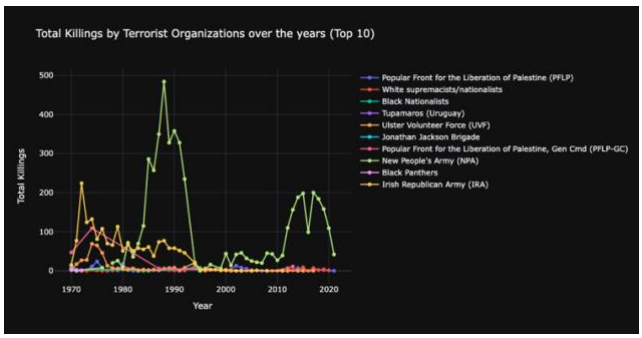


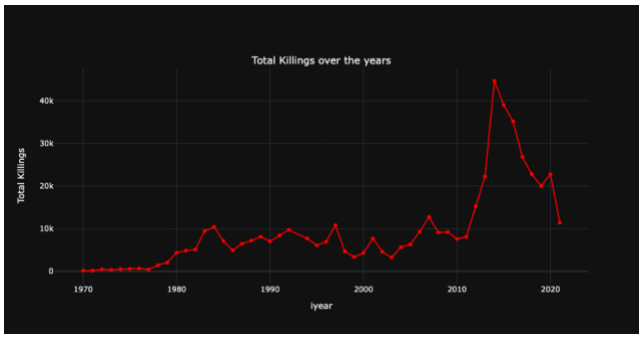
4. PROPOSED WORK

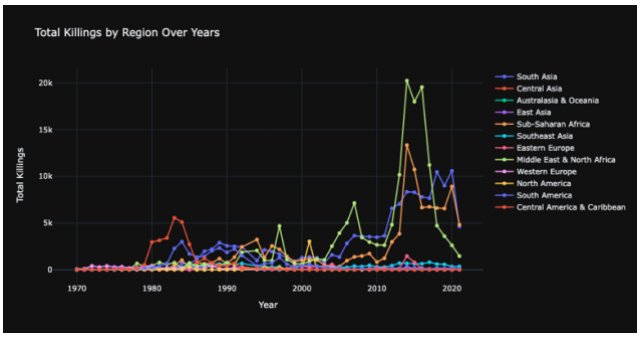
I. Exploratory Data Analysis

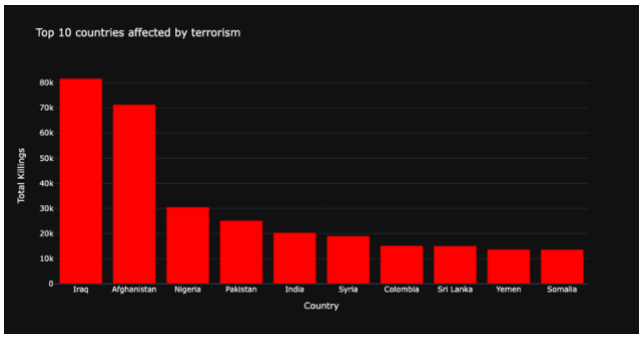
Shape of Data: There are 214666 records and 135 attributes.
Dataset Information: GTD dataset has 10 numerical features and 125 categorical features.
Duplicate Values: There are no duplicate values.
Missing Values: There are missing values.

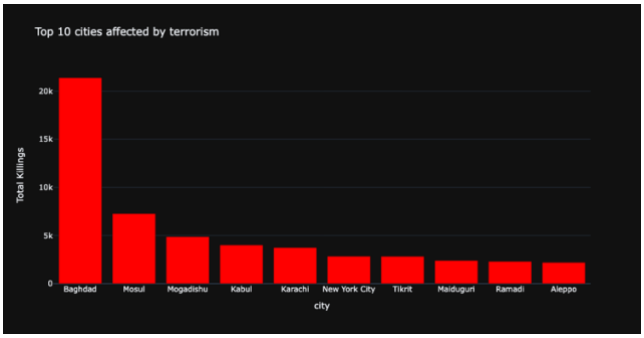


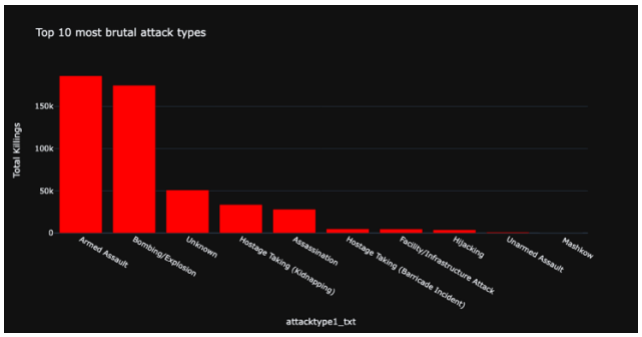


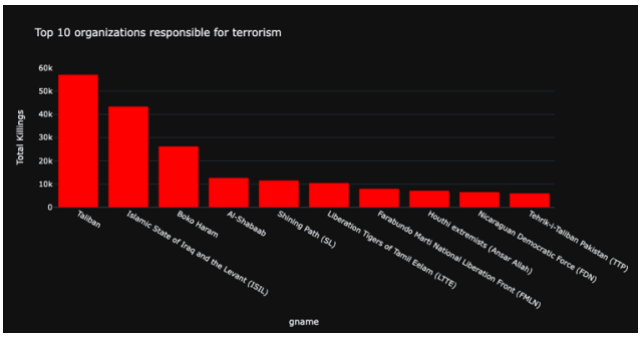

Total Killings over the years
Total Killings
iyear

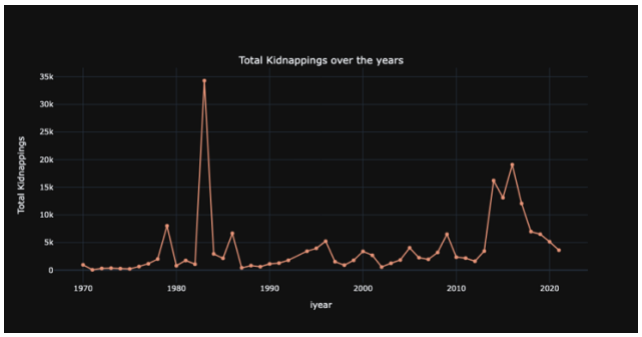

Total Killings by Region Over Years
Total Killings
Year
South Asia
Central Asia
Australasia & Oceania
East Asia
Sub-Saharan Africa
Southeast Asia
Eastern Europe
Middle East & North Africa
Western Europe
North America
South America
Central America & Caribbean

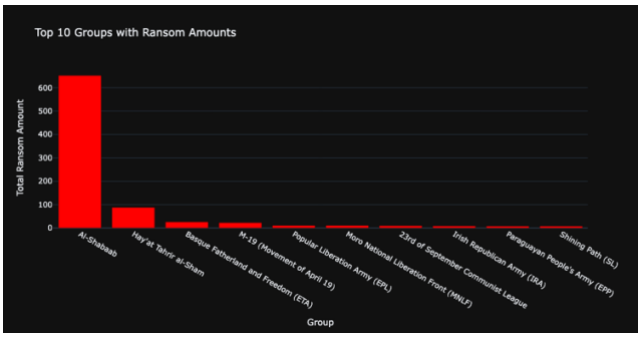

Top 10 countries affected by terrorism
Total Killings
Country
Iraq
Afghanistan
Nigeria
Pakistan
India
Syria
Colombia
Sri Lanka
Yemen
Somalia

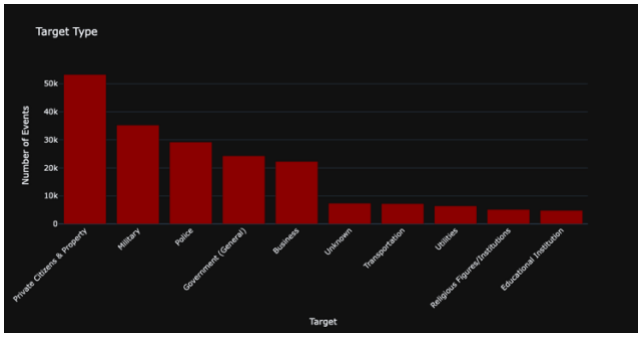

Top 10 cities affected by terrorism
Total Killings
city
Baghdad
Mosul
Mogadishu
Kabul
Karachi
New York City
Tikrit
Maiduguri
Ramadi
Aleppo

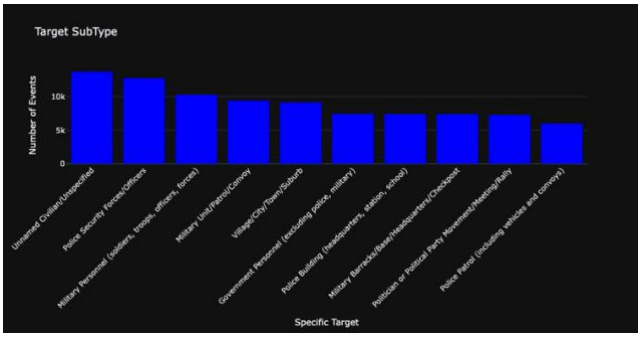

Top 10 most brutal attack types
Total Killings
attacktype1_txt
Armed Assault
Bombing/Explosion
Unknown
Hostage Taking (Kidnapping)
Assassination
Hostage Taking (Barricade Incident)
Facility/Infrastructure Attack
Hijacking
Unarmed Assault
Mashkew

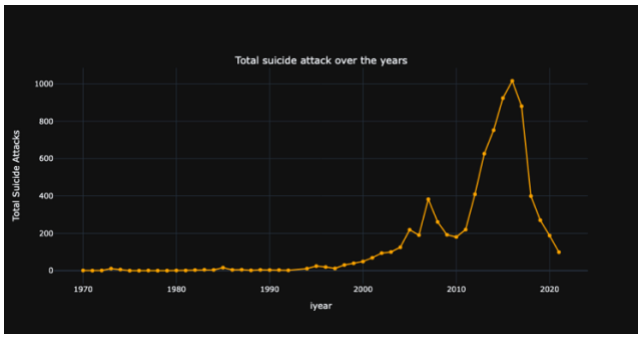

Top 10 organizations responsible for terrorism
Total Killings
gname
Taliban
Islamic State of Iraq and the Levant (ISIL)
Boko Haram
Al-Shabaab
Shining Path (SL)
Liberation Tigers of Tamil Eelam (LTTE)
Farabundo Marti National Liberation Front (FMLN)
Houthi extremists (Ansar Allah)
Nicaraguan Democratic Force (FDN)
Tehrik-i-Taliban Pakistan (TTP)

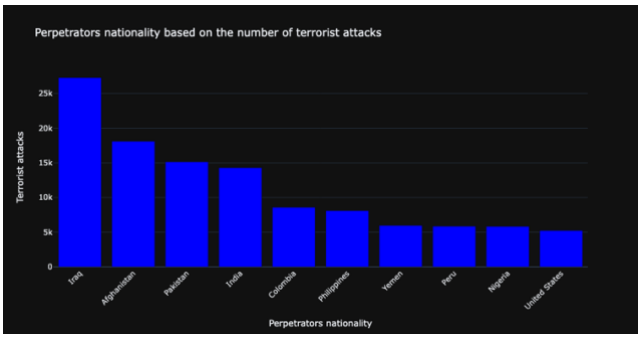

Total Kidnappings over the years
Total Kidnappings
iyear

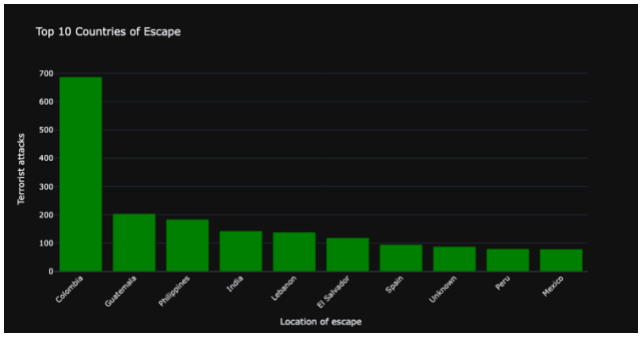

Top 10 Groups with Ransom Amounts
Total Ransom Amount
Group
Al-Shabaab
Hay'at Tahrir al-Sham
Basque Fatherland and Freedom (ETA)
M-19 (Movement of April 19)
Popular Liberation Army (EPL)
Moro National Liberation Front (MNLF)
23rd of September Communist League
Irish Republican Army (IRA)
Paraguayan People's Army (EPP)
Shining Path (SL)

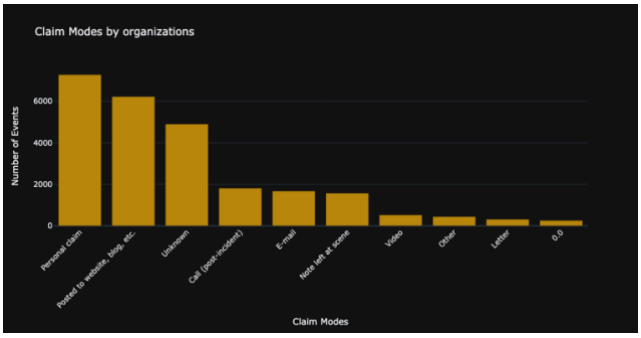

Target Type
Number of Events
Target
Private Citizens & Property
Military
Police
Government (General)
Business
Unknown
Transportation
Utilities
Religious Figures/Institutions
Educational Institution


Target SubType
Number of Events
Specific Target
Unnamed Civilian/Unspecified
Police Security Forces/Officers
Military Personnel (soldiers, troops, officers, forces)
Military Unit/Patrol/Convoy
Village/City/Town/Suburb
Government Personnel (excluding police, military)
Police Building (headquarters, station, school)
Military Barracks/Base/Headquarters/Checkpost
Politician or Political Party Movement/Meeting/Rally
Police Patrol (including vehicles and convoys)

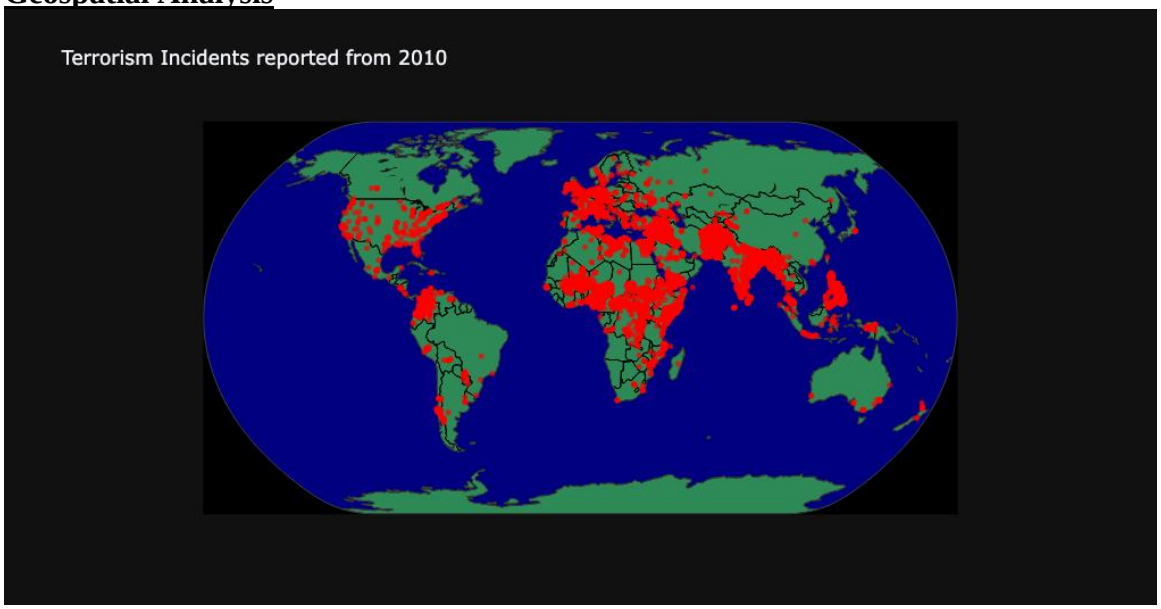








Geospatial Analysis



Network Analysis

Nodes: 3764, Edges: 4344

Number of communities detected: 10

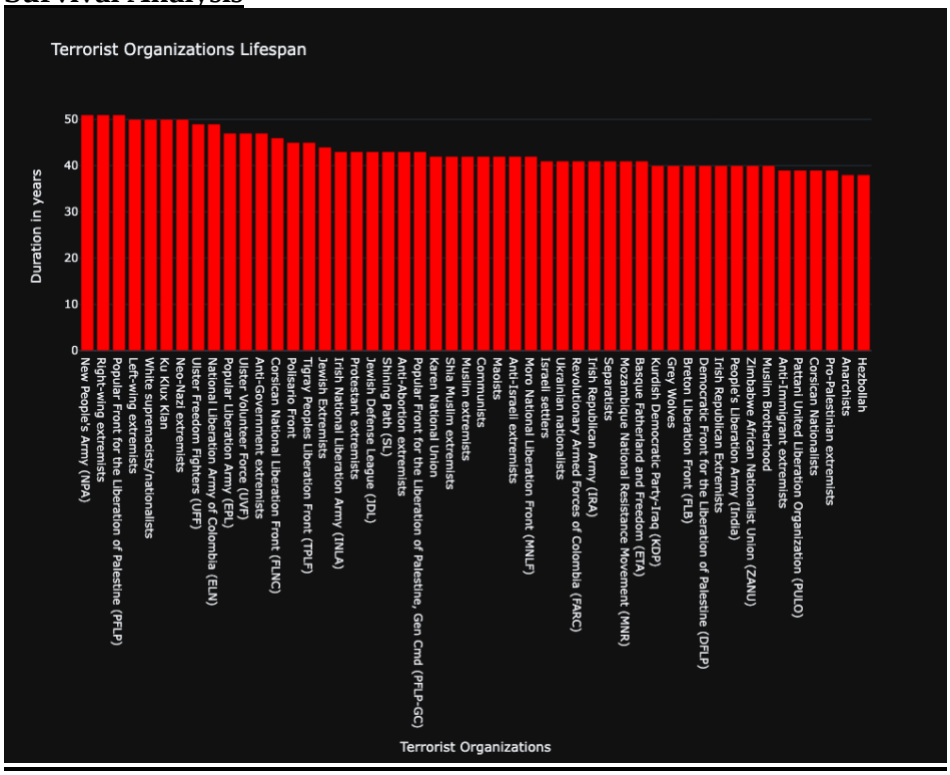
Top 5 central nodes:

{'Western Europe': 0.3765424942777016, 'Middle East & North Africa': 0.3377507908135361, 'South Asia': 0.2695946407685367, 'Sub-Saharan Africa': 0.24484092040275482, 'South America': 0.1598010546307253}

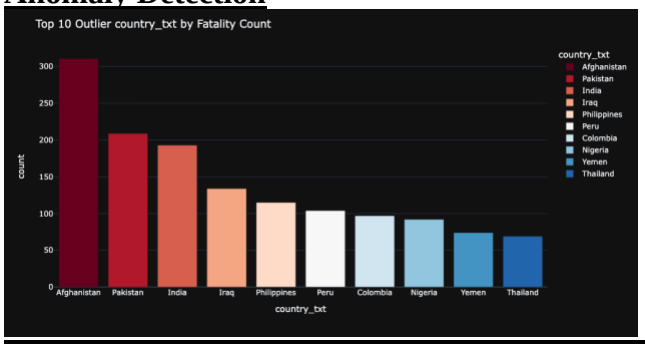
Top 5 nodes by Closeness Centrality:

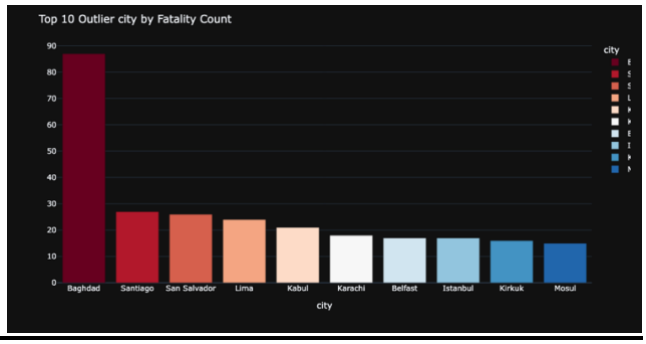
{'Separatists': 0.4722640562248996, 'Anti-Government Group': 0.46664186507936506, 'Students': 0.4596872709504031, 'Guerrillas': 0.45946275946275944, 'Rebels': 0.4414594087282966}

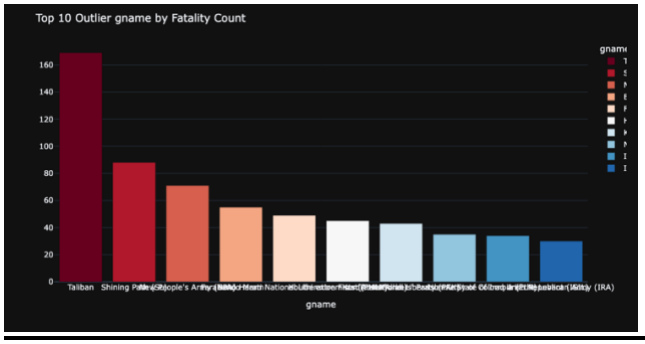
Survival Analysis

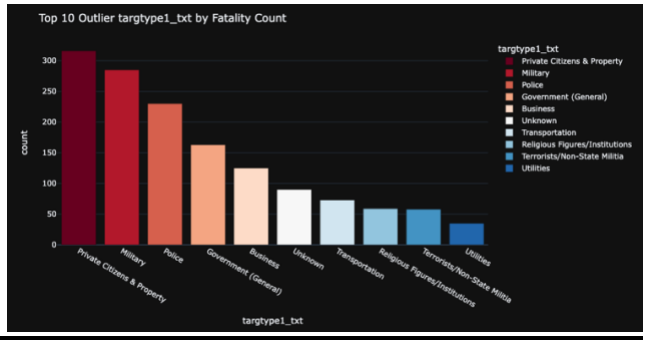


Anomaly Detection

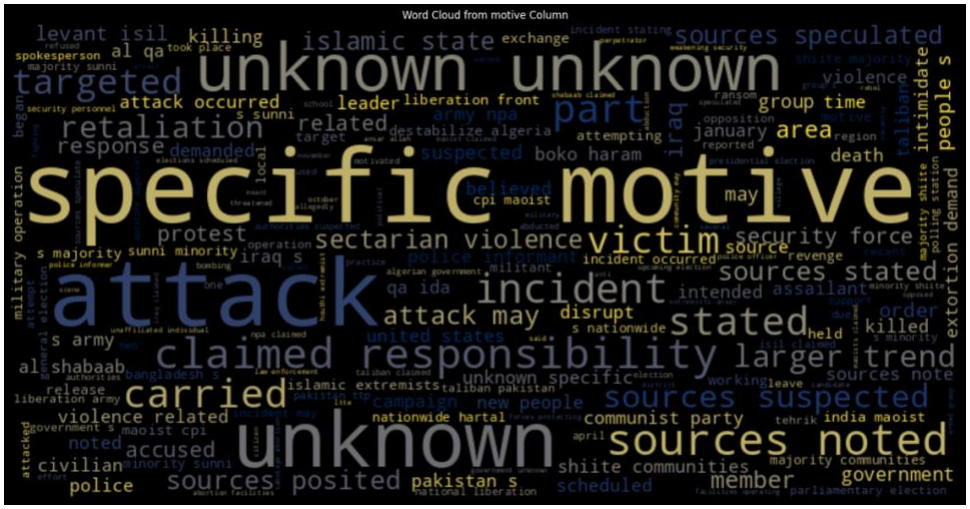


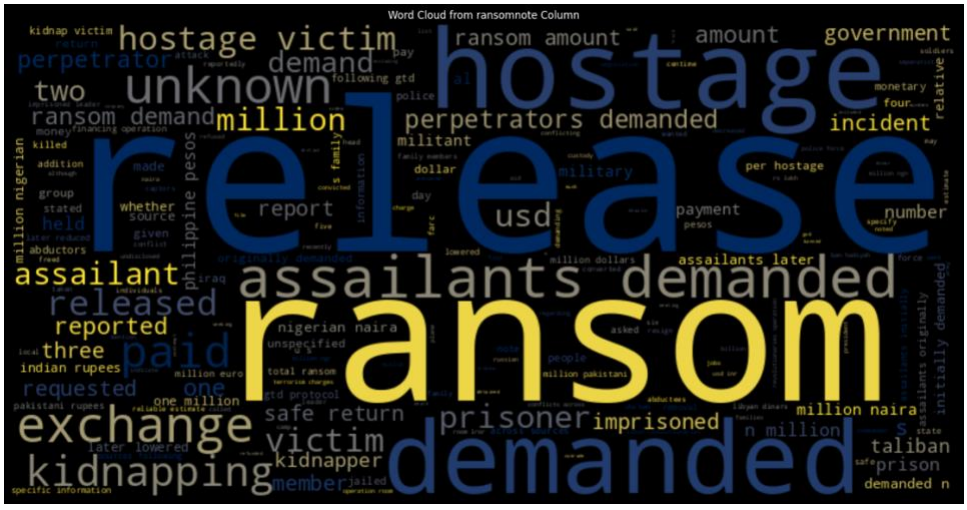






Word Cloud





II. Feature Engineering

Data Cleaning: The categorical features were cleaned by removing special characters, eliminating multiple spaces, and converting all text to lowercase.

Text Processing: The text was split into individual words, removed stopwords, and transformed into vectors using CountVectorizer.

III. Topic Modeling

This code evaluated different numbers of topics for LDA and calculated how well each model fit the data (perplexity), and then identified the best model with the optimal number of topics based on the lowest perplexity score.

Topic Modeling using LDA

- num_topics: [5, 10, 15, 20]

Best Parameters:
num_topics = 5, logPerplexity: 6.920436509164585

Distribution of words most important for each topic
Topic 0 –
Top Words: general, electric, belonging, unknown, damaged, strike, part, two, bombed, nearby
Topic 1 –
Top Words: city, sioux, attacks, south, poles, many, nebraska, beef, packers, iowa
Topic 2 –
Top Words: ghitani, grisha, horses, niyaz, kuyi, paika, morgado, maraqisha, calamba, doberdan
Topic 3 –
Top Words: police, states, united, unknown, perpetrators, casualties, colorado, car, building, bomb
Topic 4 –
Top Words: detonated, company, san, kresge, exploded, inflicted, francisco, oakland, paint, juan

IV. Data Modeling

A predictive model was developed to forecast the likelihood of future terrorist incidents across regions, countries, states, and cities by analyzing attack, target, and weapon types, as well as the nationality of terrorists and organizations.

The data was randomly split into a training set (80%) and a test set (20%).

The following machine learning model was trained on the training set using cross-validation with 10 folds and evaluated on the test set based on following parameters:

- XGBoost Model

```
param_grid = ParamGridBuilder() \
        .addGrid(xgboost.max_depth, [1, 2]) \
        .addGrid(xgboost.n_estimators, [10, 20]) \
        .addGrid(xgboost.learning_rate, [0.01, 0.1]) \
        .build()
```

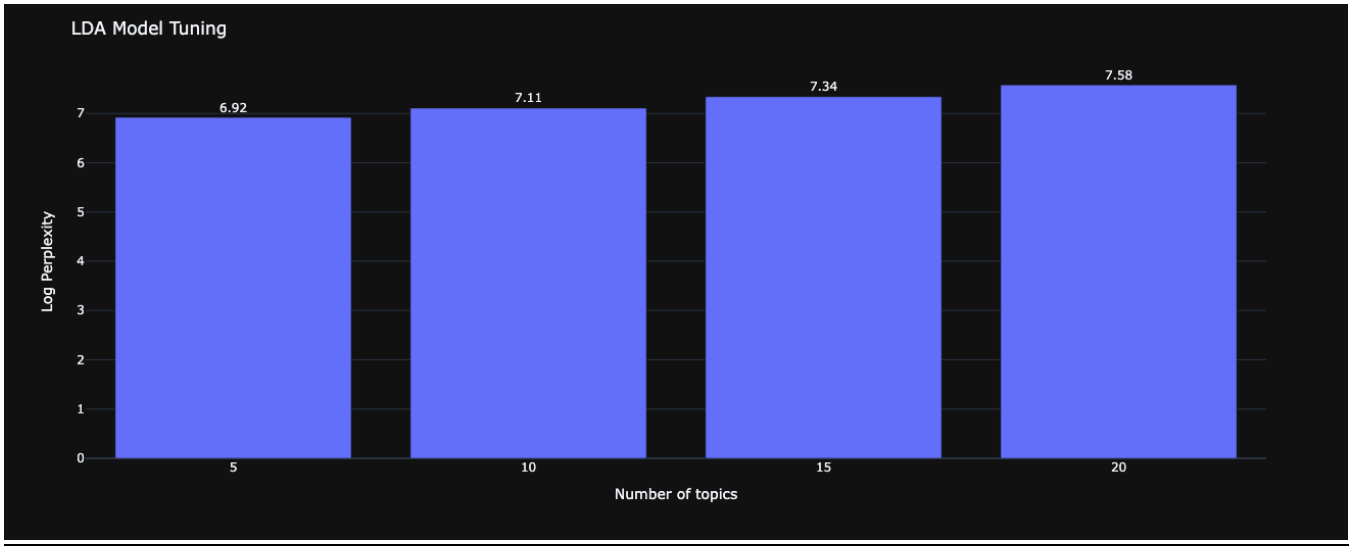
- XGBoost Model (With Political Stabilty)

```
param_grid = ParamGridBuilder() \
        .addGrid(xgboost.max_depth, [1, 2]) \
        .addGrid(xgboost.n_estimators, [10, 20]) \
        .addGrid(xgboost.learning_rate, [0.01, 0.1]) \
        .build()
```

5. RESULTS

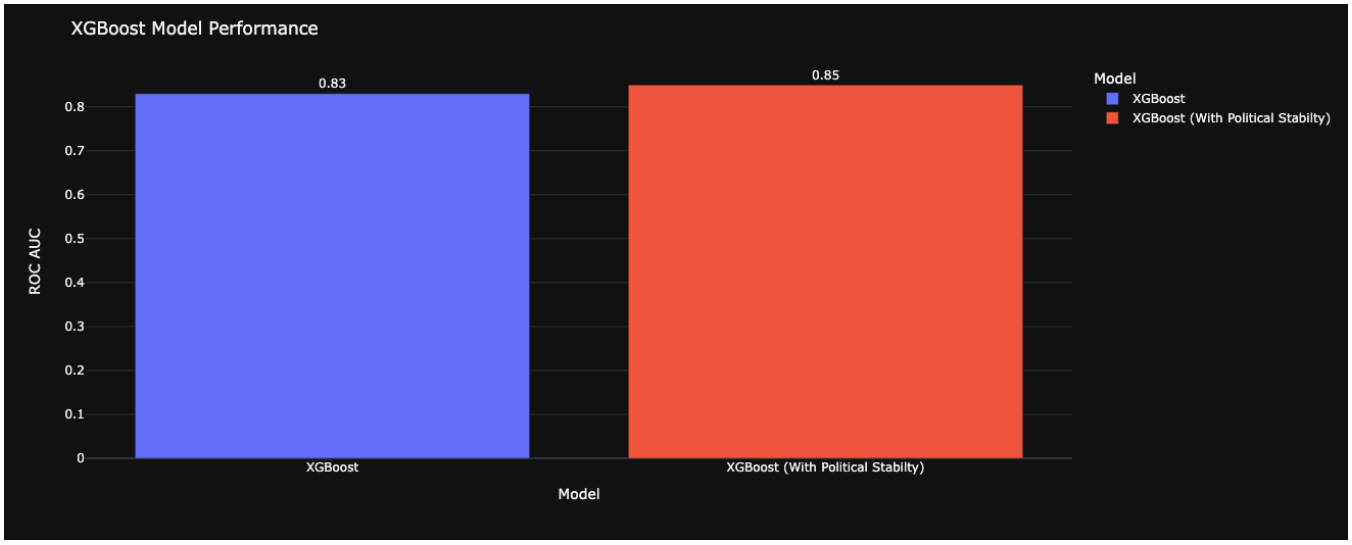
- Topic Modeling (Using LDA)

Number of topics	Log Perplexity
5	6.92
10	7.11
15	7.34
20	7.58



- XGBoost Model Performance

Model	ROC AUC
XGBoost	0.83
XGBoost (With Political Stabilty)	0.85



6. CONCLUSION

- Killings from terrorist activities increased significantly from 2010-2020, reaching a peak of 44,640 deaths in 2014 globally.
- Central America and the Caribbean saw an initial spike in terrorist killings to 5,543 deaths in 1983.

- The Middle East and North Africa hit a high of 20,000 deaths between 2012-2018, after a prior peak of 7,000 in 2007.
- South Asia and Sub-Saharan Africa reached peak death tolls recently of 10,000 in 2018-2020 and 13,000 in 2014, respectively.
- Iraq suffered the highest number of killings with 81,000 deaths. Afghanistan followed with 71,000 deaths.
- Baghdad, Iraq was the city most impacted by terrorism with 21,000 killings.
- The Taliban, ISIL and Boko Haram have been responsible for the most killings over time, resulting in 57,000, 43,000 and 26,000 deaths respectively.
- Armed assaults and bombings were the most lethal attack types employed, each resulting in over 170,000 deaths.
- Kidnappings spiked in 1984 with 34,000 kidnappings globally, with a recent resurgence to 19,000 in 2016. Al-Shabaab demands the highest average ransom amounts.
- Suicide attacks have increased over time, peaking at 1,016 attacks in 2016.
- Iraq contributed the most attacks with 27,000. Afghanistan and Pakistan follow with 18,000 and 15,000 attacks respectively.
- Colombia is the most common country for terrorists to escape to after attacks.
- The optimal LDA model had 5 topics and achieved the least log perplexity score of 6.92.
- Based on ROC AUC, the XGBoost model with political stability performed better, achieving a score of 0.85 as compared to base XGBoost Model with a score of 0.83.

7. REFERENCES

[1] https://spark.apache.org/docs/latest/api/python/index.html
[2] https://www.start.umd.edu/gtd/
[3] https://databank.worldbank.org/