Advanced Pattern Recognition

CS6103

Assignment 1

Project Report

Global Terrorism Data Analysis & Prediction

Name: Akanksha Kumari

Roll No: 2511AI23

Date: 19/08/2025

Introduction

Terrorism is a major global concern that affects human lives, social stability, and economic growth. Over the years, terrorist activities have become more frequent, more violent, and more widespread across the globe. Governments and security agencies require accurate data-driven insights to identify patterns, anticipate risks, and develop effective preventive measures.

This project uses the Global Terrorism Database (GTD), which contains more than 180,000 incidents from 1970 to 2017, to analyze and predict patterns in terrorism. Our work focuses on understanding where, when, and how terrorism occurs and whether we can predict the success of an attack using machine learning techniques.

Key objectives of the project:

- Trend Analysis: Study the yearly and regional rise in terrorist incidents.
- Attack Profiling: Identify the most common attack types, targets, and terrorist groups.
- Outcome Prediction: Build a Logistic Regression model to predict whether an attack will be successful.
- Feature Insights: Use Principal Component Analysis (PCA) to reduce dimensionality and visualize patterns in the data.

By combining Exploratory Data Analysis (EDA) with Machine Learning, this project aims to uncover meaningful insights and provide a foundation for future predictive modeling efforts in terrorism research.

Dataset Description

- **Source:** Global Terrorism Database (University of Maryland)
- **Coverage:** 1970–2017
- **Total Records:** ~180,000 incidents

Column	Description
Year	Year of attack
Country	Country where attack occurred
Region	Region of attack
AttackType	General type of attack
Target	Victim or target of the attack
Weapon_type	Weapon used
Group	Terrorist group responsible
Killed / Wounded	Number of casualties
success	Target variable (1 = successful, 0 = failed)

Target Variable: success (binary classification problem)

Methodology

The project followed a systematic approach:

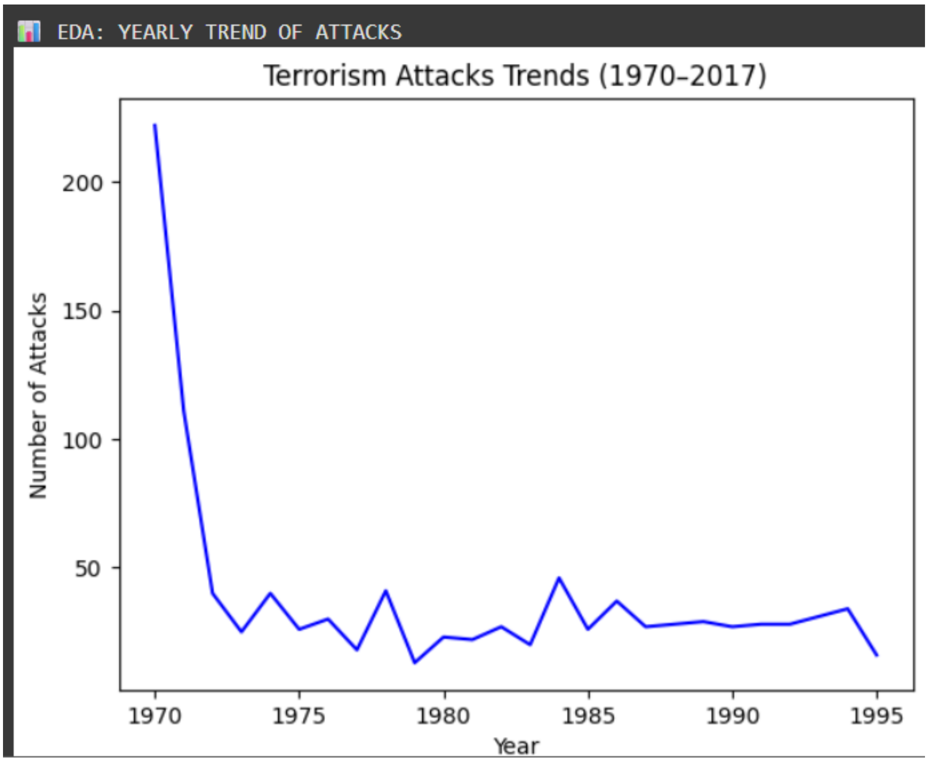
1. **Data Preprocessing**
    - Renamed columns for readability.
    - Selected relevant features and dropped missing values.
    - Converted categorical variables into numeric format for modeling.
2. **Exploratory Data Analysis (EDA)**
    - Visualized global and regional terrorism trends.
    - Found top affected countries and groups.
    - Studied attack type distribution and success rates.
    - Generated a choropleth map to visualize country-wise attack intensity.
3. **Machine Learning (ML) Workflow**
    - **Model Used:** Logistic Regression (binary classification).
    - **Data Split:** 80% training, 20% testing.
    - **Evaluation Metrics:** Accuracy, Confusion Matrix, Precision, Recall, F1-Score.
    - **Dimensionality Reduction:** PCA applied to visualize data in 2D space.

Implementation and Results

We performed a detailed EDA to understand trends, patterns, and distributions.

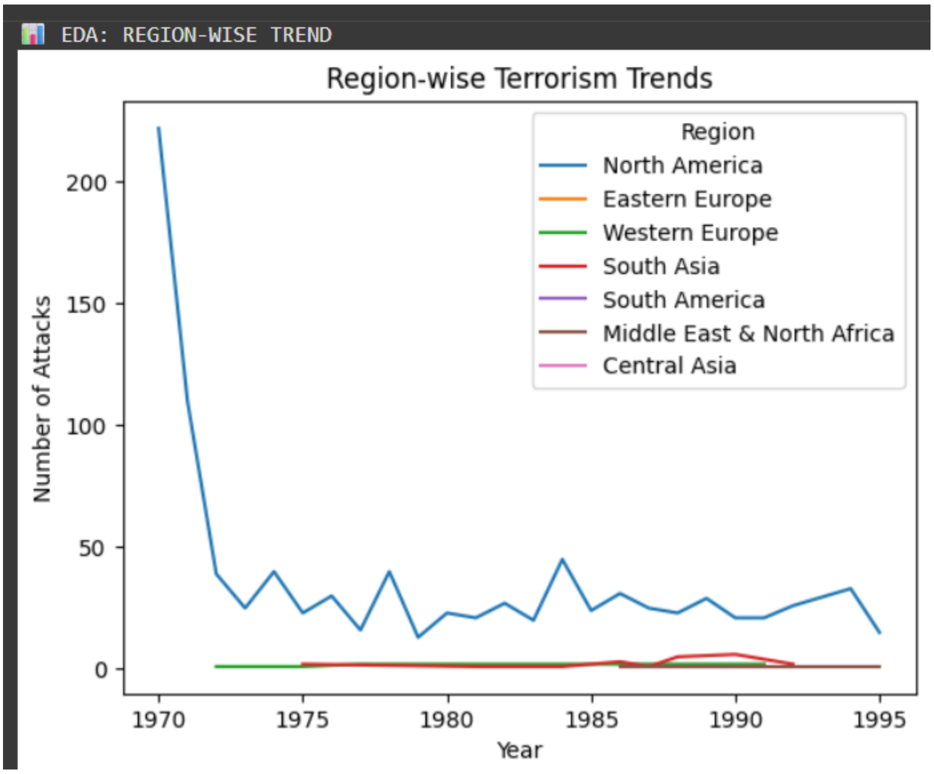
Global Attack Trends (Yearly):

- Shows the number of attacks per year from 1970–2017.
- A sharp rise is observed after 2010, with a peak in 2014.



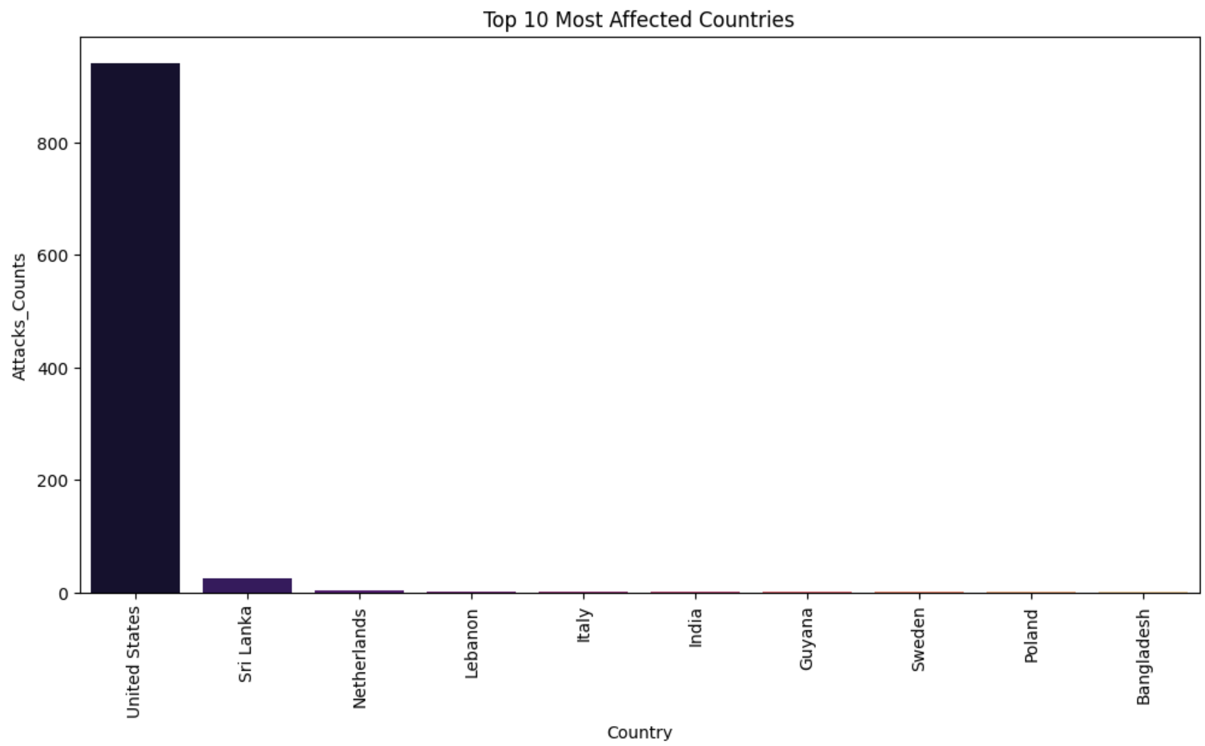
Region-Wise Distribution:

- Highlights which regions have faced the most terrorist attacks.
- South Asia and the Middle East & North Africa contribute the highest number of incidents.



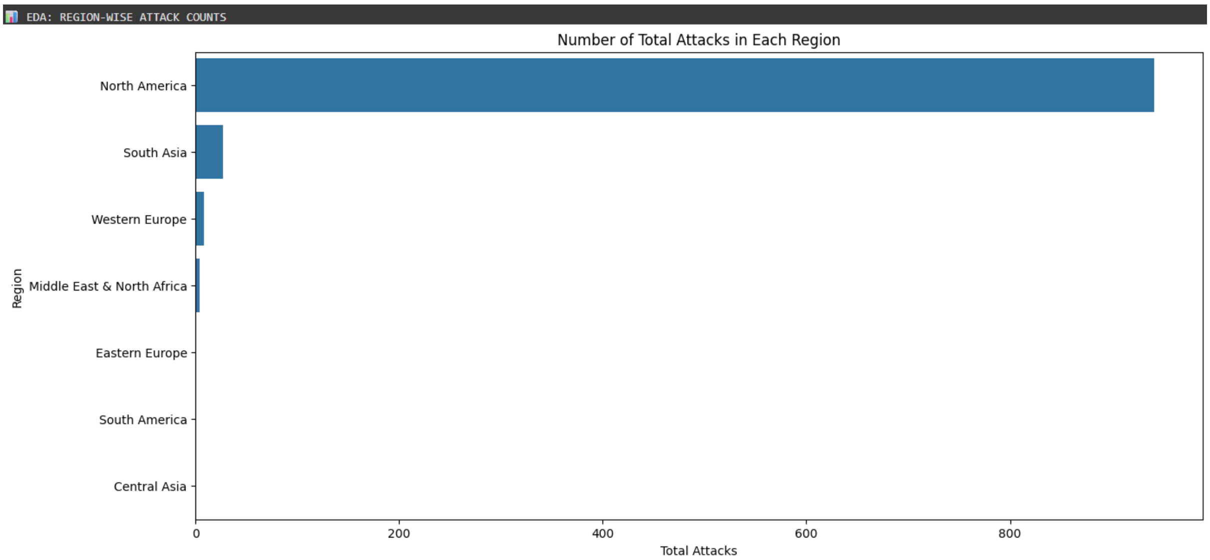
Top 10 Affected Countries:

- Lists countries with the highest number of attacks.
- Iraq, India, Pakistan, and Afghanistan dominate the list.



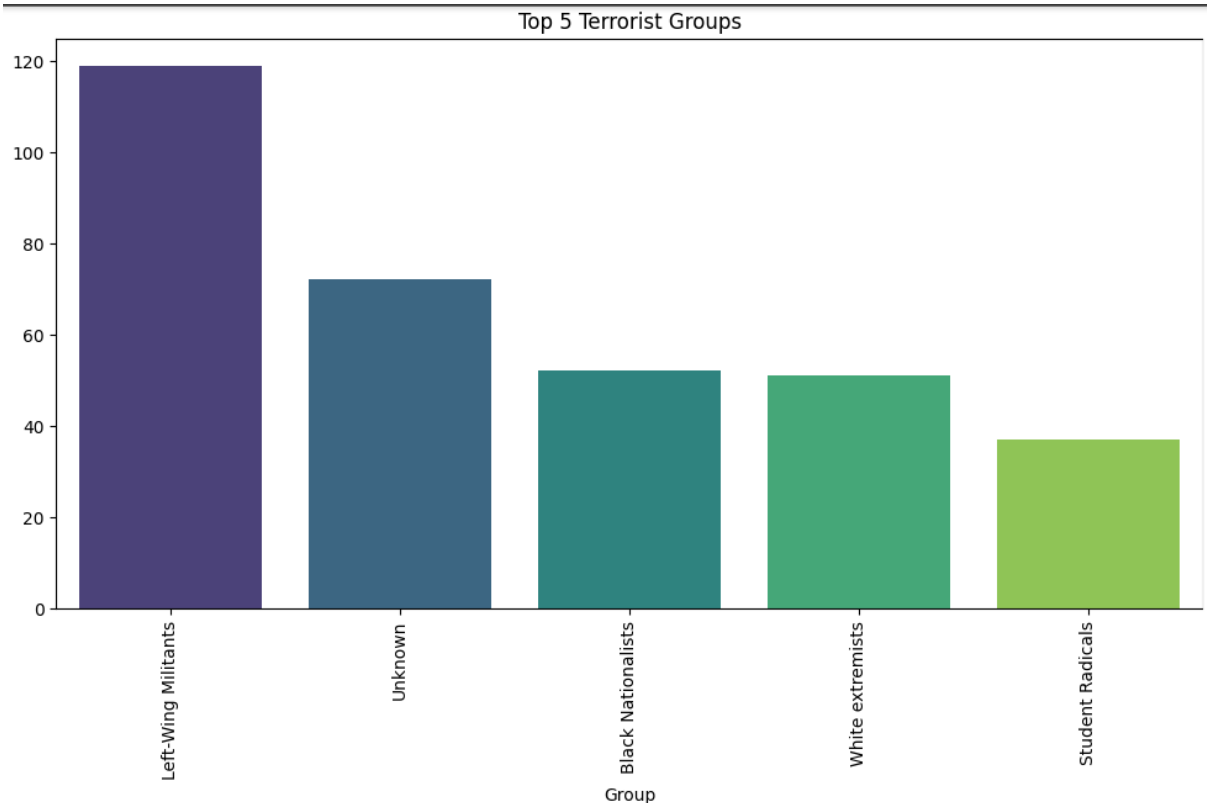
Number of attacks in each region:

- North America has the most attacks significantly higher than others.
- South America, Central Asia and Eastern Europe has very less attacks.



Top 5 Terrorist Groups:

- Identifies the most active terrorist groups worldwide.
- Left wing militants is the most active, followed by ISIS and others.

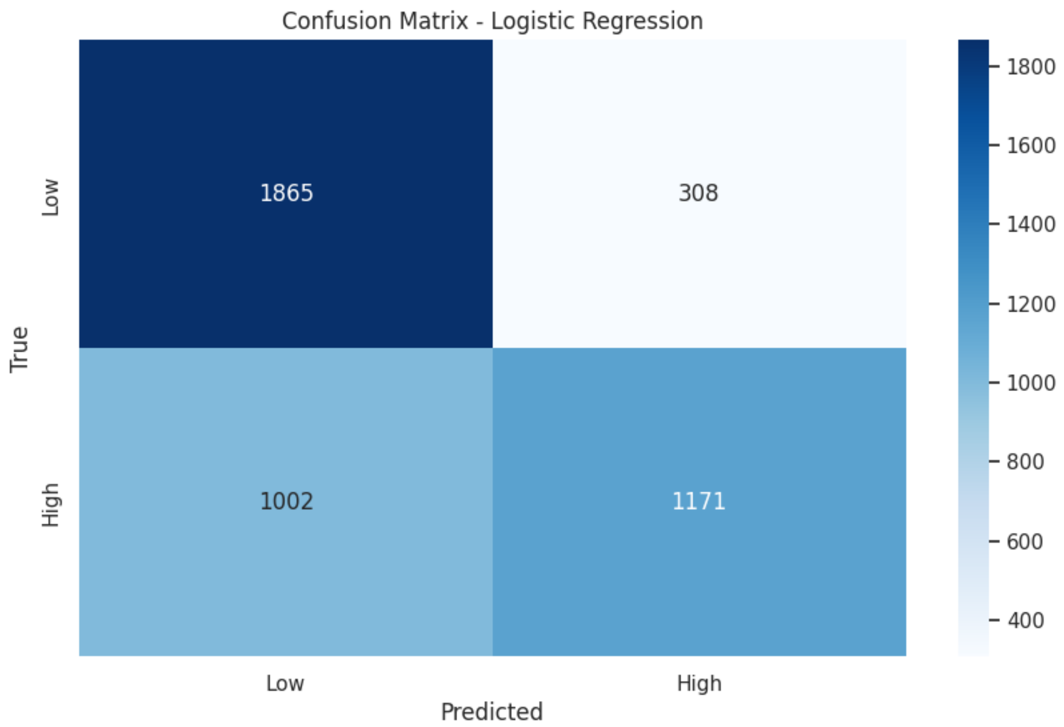


Classification report

Provides detailed metrics:

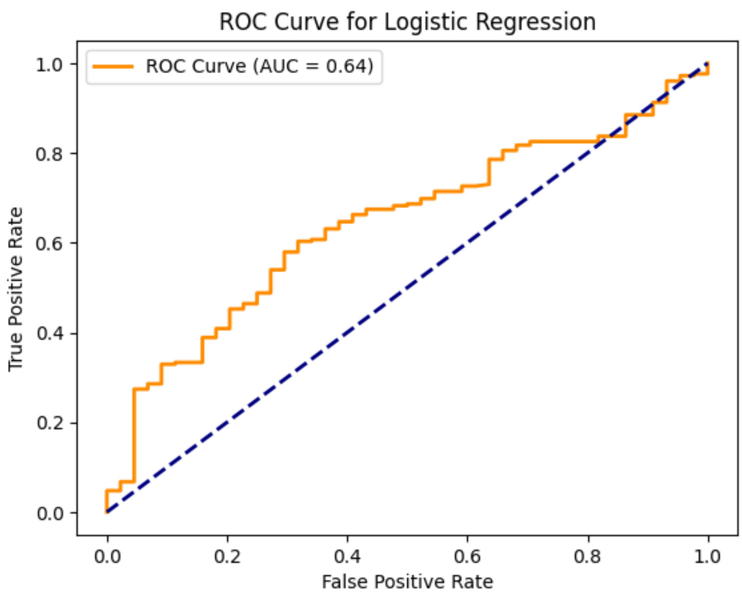
- **Precision:** Fraction of predicted successes that were correct
- **Recall:** Fraction of actual successes detected
- **F1-Score:** Balance between precision and recall
- **Confusion matrix** visualizes model predictions vs. actual outcomes. Shows how any successes and failures were correctly and incorrectly classified.

```
🤖 Training Logistic Regression Model...

📌 MODEL EVALUATION
Accuracy: 0.8514

Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        44
           1       0.85      1.00      0.92       252

    accuracy                           0.85       296
   macro avg       0.43      0.50      0.46       296
weighted avg       0.72      0.85      0.78       296


Confusion Matrix:
 [[  0  44]
 [  0 252]]
```



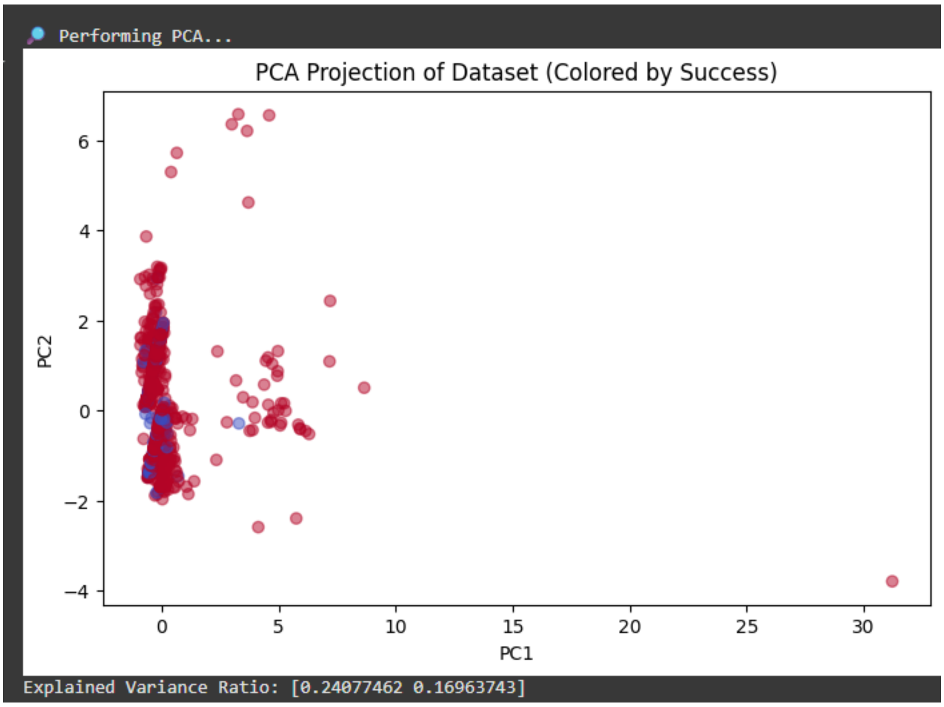
ROC Curve & AUC

- Evaluates model discrimination ability.
- Higher AUC indicates better separation between success and failure classes.



PCA Visualization (2D Scatter)

- Plots first two principal components for a 2D projection of the dataset.
- Points are colored by attack success (0 = Failure, 1 = Success).
- Helps visualize clustering and separability of classes.



Predictions on 10 Random Samples

- Compares actual vs predicted success for 10 random events.
- Includes predicted probability to show model confidence.

```
🎲 Random Sample Predictions (10 rows):
       Year     Month       Day   Country   Region  AttackType    Killed
0 -0.936363  0.799615 -1.464565  0.171491 -0.16062   -0.259676 -0.136661
1 -0.443580  0.799615 -1.127721  0.171491 -0.16062   -1.078653 -0.136661
2  1.897141 -1.217875 -0.004907  0.171491 -0.16062    2.197255 -0.136661
3 -0.443580  1.087828 -0.229469  0.171491 -0.16062   -0.259676 -0.136661
4 -1.059559 -0.353236 -1.576847  0.171491 -0.16062   -0.259676 -0.136661
5 -1.059559 -0.929662 -0.454032  0.171491 -0.16062    0.559301 -0.136661
6  1.404357 -1.506088  1.679315  0.171491 -0.16062   -1.078653  0.108283
7  1.404357  0.223189  1.117908 -2.058646  4.35509   -1.897629  2.067833
8 -1.059559 -0.353236 -1.689128  0.171491 -0.16062    0.559301 -0.136661
9 -1.059559 -0.929662  1.791596  0.171491 -0.16062   -0.259676 -0.136661

    Wounded  Weapon_type  Actual Success  Predicted Success  \
0 -0.139478    -0.931401               0                  1
1 -0.139478    -0.062682               0                  1
2 -0.139478    -0.062682               1                  1
3 -0.139478    -0.931401               1                  1
4 -0.139478    -0.931401               1                  1
5 -0.139478     0.806037               1                  1
6 -0.139478     1.674756               1                  1
7 -0.139478     1.674756               1                  1
8 -0.139478     0.806037               1                  1
9 -0.139478    -0.931401               0                  1

   Predicted Probability
0               0.776914
1               0.852899
2               0.815377
3               0.778908
4               0.749230
5               0.898858
6               0.947001
7               0.998806
8               0.906120
9               0.732529
```

Conclusion

- Terrorism has increased significantly in the last two decades.
- Bombings and explosions are the most frequent attack type.
- Certain regions are highly vulnerable, particularly South Asia and MENA.
- Logistic Regression model is effective in predicting attack outcomes, meaning historical patterns are strong predictors.
- PCA shows that data points for successful and failed attacks form separable clusters.

The analysis demonstrates that terrorism has a clear geographic and temporal concentration. Machine learning models like Logistic Regression can help estimate attack outcomes, which may be useful for risk assessment. Future work could involve:

- Clustering countries by attack characteristics
- Time series forecasting of attack frequency
- More advanced models like Random Forest or XGBoost

References

- **Dataset:** Global Terrorism Database (GTD), University of Maryland
- **Libraries Used:** pandas, numpy, seaborn, matplotlib, scikit-learn, plotly