Course Project Report: Quantitative Analysis of Terrorism Incidents and Their Impact on Human Casualties

STA 2101: Statistics & Probability

Student Name: Raian Zaman
Student ID: 222014031
University of Liberal Arts Bangladesh (ULAB)

December 24, 2025

Abstract

This project uses basic statistical and probability ideas to examine a large dataset on terrorist incidents. It focuses on understanding how attacks happen, how serious their impacts are, and how different features of an incident relate to the number of people killed. Throughout the different stages, several methods are applied. These include comparing sampling methods, creating frequency distributions, summarizing key numerical variables, defining probability events, and conducting simple linear regression analysis with `nkill` as the main response variable. The results show that the number of recorded attacks has risen over the years, while most incidents lead to little or no loss of life. However, a small number of extreme attacks cause a significant portion of total casualties. The probability analysis, which includes Bayes' rule, indicates a clear link between the success of an attack and whether it leads to fatalities. The regression analysis also reveals that the number of injured individuals has a moderate connection to fatalities. Other individual variables are less informative, suggesting that more detailed models are necessary to fully explain the patterns seen in the data.

Contents

1 Milestone 1: Dataset Selection

- **Dataset Name:** Global Terrorism Database (GTD)
- **Dataset URL:** https://www.kaggle.com/datasets/START-UMD/gtd
- **Description:** The Global Terrorism Database (GTD) is an extensive, publicly accessible dataset that records terrorist attacks that took place globally between 1970 and 2017. A single recorded episode is represented by each row in the dataset, which contains comprehensive details about the event's location, timing, attack characteristics, and outcomes.

  The dataset contains variables describing the **type of attack**, **weapons used**, **target category**, and whether the attack was considered successful. In addition, several numerical variables capture the human impact of terrorism, including the **number of fatalities** (`nkill`) and the **number of injured individuals** (`nwound`). These variables allow meaningful quantitative analysis of attack severity and outcomes.

  The GTD is ideal for statistical analysis because of its scale, worldwide reach, and combination of numerical and category information. It supports a variety of approaches utilized in this course, including as probability analysis, descriptive statistics, sampling procedures, and basic linear regression. Most significantly, the dataset offers practical context, which makes the statistical results applicable for comprehending trends and dangers related to terrorist attacks.

2 Milestone 2 – Probability Sampling Analysis

2.1 Column Selection Justification

2.1.1 Numeric Column: `nkill` (Number of Individuals Killed)

Significance: `nkill` records the number of fatalities in a terrorist attack and is a key quantitative indicator of severity and human cost. It allows direct numerical comparisons of violence across events, regions, and time.

Analytical role: I use `nkill` as the primary numeric variable to compute means, medians, and standard deviations for the population and each sample. These statistics form the basis for evaluating how well different sampling methods approximate the population.

2.1.2 Stratified Sampling Column: `region_txt`

Rationale: `region_txt` groups attacks into broad geographic regions (e.g., Middle East & North Africa, South Asia, Western Europe). Terrorism patterns and impacts differ by

region due to geopolitical and socio-economic factors.

Benefits: Stratifying on `region_txt` ensures that each major region appears in the sample in proportion to its presence in the population. This reduces sampling error and yields more precise estimates of parameters such as the mean of `nkill` by accounting for regional heterogeneity, rather than letting one region dominate or be under-represented by chance.

2.1.3 Cluster Sampling Column: `country_txt`

Rationale: `country_txt` naturally groups attacks by country. Incidents within a country often share causes, actors, or security contexts, so countries form realistic clusters. In practice, it is often easier and more cost-effective to sample at the cluster (country) level than to sample individual attacks spread across many countries.

Implications: When using `country_txt` as the cluster variable, the sampling unit becomes the country. Once a country is selected, all or a subset of its attacks enter the sample. This can introduce higher sampling error if clusters differ strongly from each other, but it reflects how large-scale data collection is frequently done. The chosen strategy (Option A) selects a few countries entirely and then samples from others to reach the desired total, illustrating a realistic cluster sampling design.

2.2 Comparison and Reflection

2.2.1 Summary of Key Statistics

Key statistics for the population and each sampling method (from `comparison_df`) are:

Method	Sample Size	Mean	Median	StdDev
Population	171378	2.403272	0.0	11.545741
Simple Random	50	2.280000	0.0	5.245173
Systematic	50	1.400000	0.5	2.338672
Stratified	50	1.100000	0.0	2.314550
Cluster	50	0.780000	0.0	2.375814

The population mean of `nkill` is about 2.40, while the median is 0, showing a highly skewed distribution with many events having no fatalities. The large standard deviation (11.55) reflects a wide range and a few very high-casualty attacks.

- **Simple Random Sampling (SRS)**: Mean = 2.28, close to the population mean of 2.40, with the same median (0). Its standard deviation (5.25) is much lower than 11.55, suggesting the sample likely missed some extreme, high-fatality events.

- **Systematic Sampling**: Mean = 1.40, clearly below the population mean, and median = 0.5, slightly above the population median. The standard deviation (2.34) is far lower than the population's, indicating a compressed spread.
- **Stratified Sampling**: Mean = 1.10, even further from the population mean than the systematic sample, with median = 0 and standard deviation = 2.31. It behaves similarly to the systematic sample in terms of spread.
- **Cluster Sampling**: Mean = 0.78, the lowest among the methods and a strong underestimate of the population mean. The median remains 0 and the standard deviation (2.38) is again much lower than 11.55.

In this run, SRS provides the closest estimate of the population mean. All methods, however, substantially underestimate the population standard deviation, which is expected with a small sample size ($n = 50$) and a highly skewed distribution where rare, extreme events drive much of the variance.

2.2.2 Population and Method Behavior

Population characteristics: `nkill` is strongly right-skewed, with many attacks causing no deaths and a few causing very high fatalities. The combination of mean ≈ 2.40, median = 0, and standard deviation ≈ 11.55 captures this pattern.

Mean approximation:

- **SRS** comes closest to the population mean (2.28 vs. 2.40) and matches the median.
- **Systematic**, **Stratified**, and **Cluster** sampling all underestimate the mean to varying degrees (1.40, 1.10, and 0.78, respectively), with cluster sampling performing worst in this regard.

Standard deviation: All four sampling methods underestimate the true spread. None of the samples fully capture the rare high-fatality events that inflate the population standard deviation, which is a typical challenge when sampling from a heavy-tailed distribution with a small n.

2.2.3 Insights from Visualizations

The supporting plots reinforce these numerical findings:

- **Histograms**: The SRS histogram centers near the population mean, while the other methods show distributions shifted left, matching their lower means.

- **Stratified distribution comparison**: Stratified sampling preserves the proportional representation of `region_txt` between population and sample, confirming that it succeeds in matching the regional structure even though it still underestimates `nkill`'s variability.

- **Cluster-level means**: Cluster means vary strongly between countries. The particular clusters selected in this run give a low overall sample mean, illustrating how cluster choice can heavily influence the result.

2.2.4 Effectiveness and Biases of Sampling Methods

- **SRS**: Performed best for mean estimation in this example and produced a reasonable picture of the center, but has high variability with small samples from skewed data and may miss rare extremes.

- **Systematic sampling**: Gave a lower mean and much lower spread; performance may be affected by hidden ordering patterns or simple chance given the small sample size.

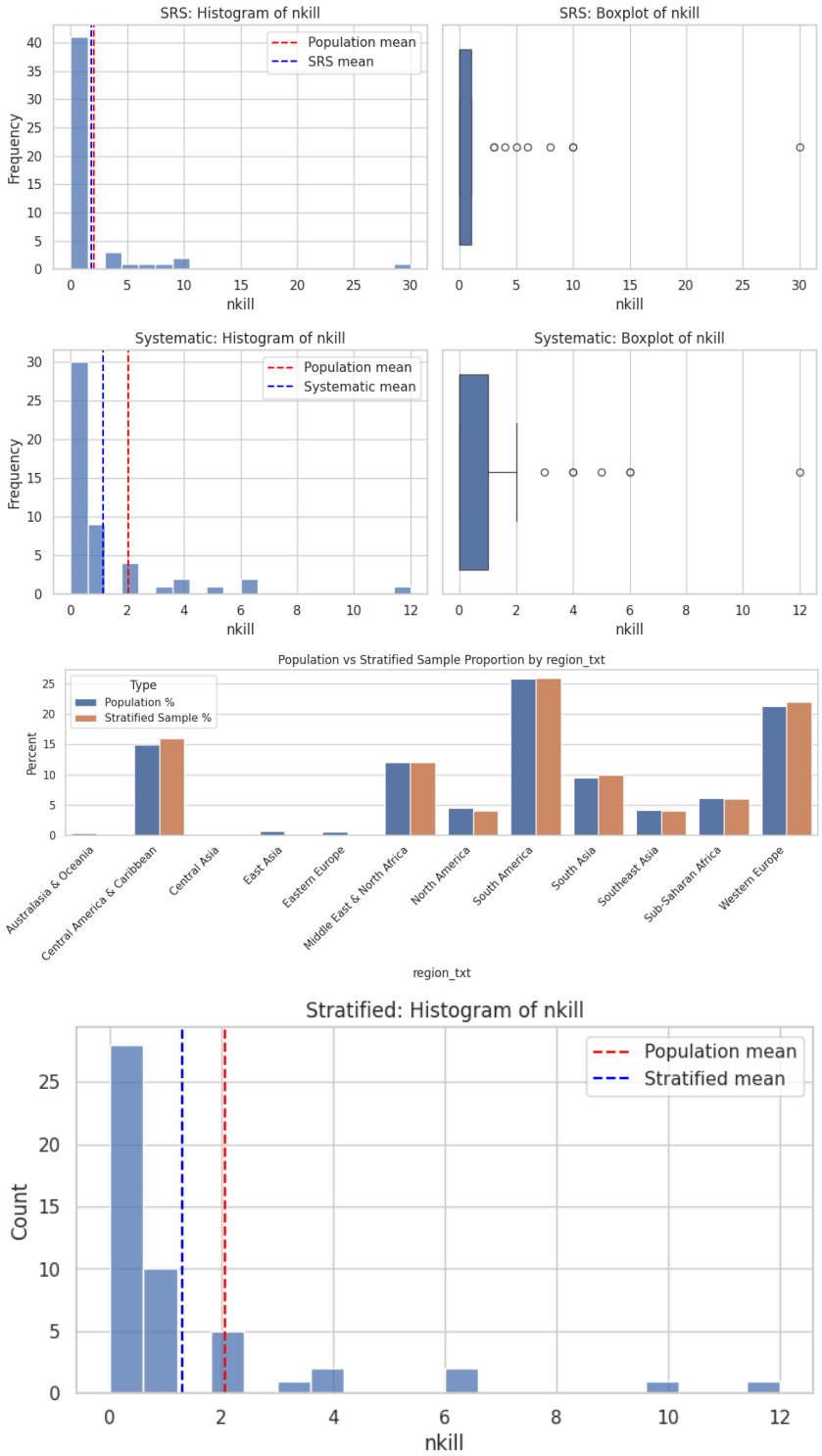
- **Stratified sampling**: Excellent at preserving the regional composition but, with few observations per stratum, struggled to capture the full variability of `nkill` within each region.

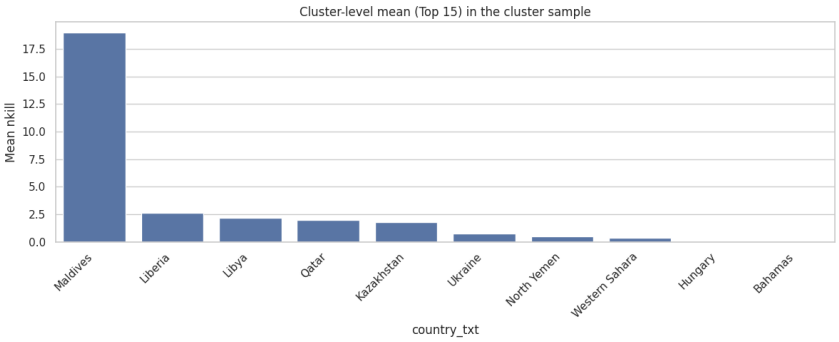
- **Cluster sampling**: Showed the strongest bias, with the lowest mean and underestimated spread. When clusters differ strongly from one another, selecting only a few countries can easily lead to samples that misrepresent the overall population.

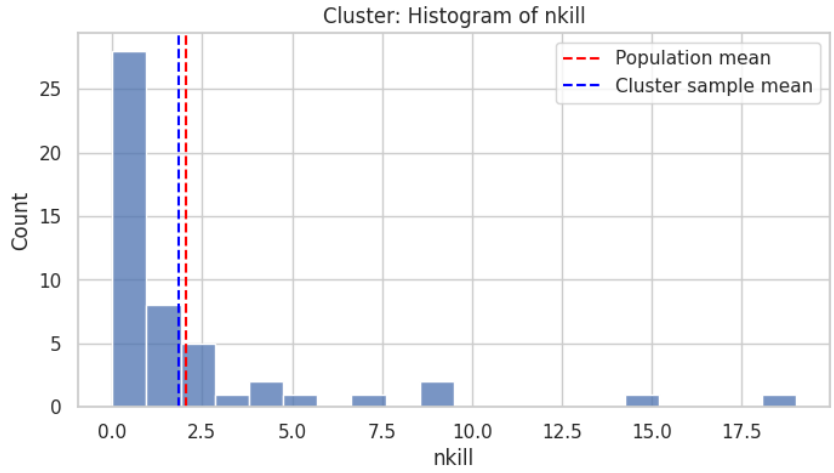
2.3 Insights and Next Steps

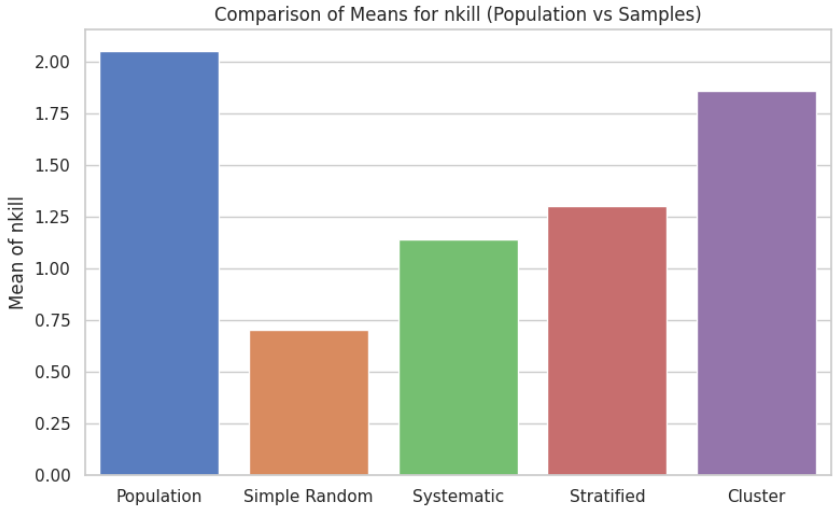
For a highly skewed variable like `nkill` with rare but extreme values, larger sample sizes are important for all methods, especially when estimating the standard deviation. If precise estimation of `nkill` is the main goal for stratified sampling, it may help to stratify on a variable more directly related to kill counts (such as `attacktype_txt` or `weaptype1_txt`). This could create strata that are more homogeneous with respect to fatalities, reduce within-stratum variance, and improve the accuracy of mean and spread estimates.

2.4 Graphical Representation









3 Milestone 3 – Frequency Distributions and Graphical Representation

3.1 Column Selection Justification

The first frequency analysis used the early numeric-like columns `eventid`, `iyear`, `imonth`, `iday`, and `extended`. `eventid` is a unique identifier, so its distribution mainly confirms uniqueness, while the others describe temporal aspects of attacks.

- `iyear` **(Year)**: Captures long-term trends in terrorism and shows how incident counts evolve over time.
- `imonth` **(Month)**: Helps detect any seasonal or monthly patterns in incident frequency.
- `iday` **(Day)**: Shows how events are spread across days of the month, even if strong patterns are less likely.
- `extended`: A binary flag (0/1) indicating whether an incident lasted beyond a single day, revealing how common prolonged attacks are.

For deeper work, other numeric or coded columns such as `nkill`, `nwound`, `country`, `region`, `attacktype1`, and `targtype1` would also be valuable for understanding impact, geography, and tactics.

3.2 Analysis and Conclusion

This milestone generated frequency tables and visualizations (histograms, frequency polygons, ogives) for `eventid`, `iyear`, `imonth`, `iday`, and `extended`.

3.2.1 `eventid`

As expected for a unique event identifier, `eventid` appears uniformly spread, with each value occurring once. The frequencies and plots mainly confirm that IDs are distinct and do not show any meaningful pattern for temporal or behavioral analysis.

3.2.2 `iyear` (Year of Attack)

The `iyear` distribution is highly informative. The histogram and frequency polygon show a strong rise in terrorist incidents over time, with a noticeable increase from the early 2000s and a peak in the 2010s. The ogive reveals a steep cumulative climb in later years, confirming that recent decades in the dataset contain many more recorded attacks.

3.2.3 `imonth` (Month of Attack)

`imonth` is spread across all 12 months. Minor fluctuations exist, but no single month consistently dominates or drops out, suggesting that—at a global level—terrorist incidents do not follow a strong monthly seasonal pattern.

3.2.4 `iday` (Day of Attack)

The distribution of `iday` is fairly uniform across the days of the month. No particular day number stands out as especially common or rare. This points to timing within the month being driven more by opportunity or context than by the calendar day itself.

3.2.5 `extended` (Extended Attack)

`extended` indicates whether an incident lasted more than one day. The frequencies and histograms show that the vast majority of attacks are non-extended (single-day), while only a small fraction are marked as extended. The ogive rises quickly with the non-extended category and then levels off, emphasizing how rare multi-day operations are compared with one-off events.

In summary, the temporal frequency analysis reveals a clear long-term increase in recorded terrorist incidents, but no strong month-of-year or day-of-month seasonality, and shows that most attacks are short, single-day events. This provides a basic temporal profile and points toward further work on why incidents have escalated over time and how extended attacks differ from standard ones.

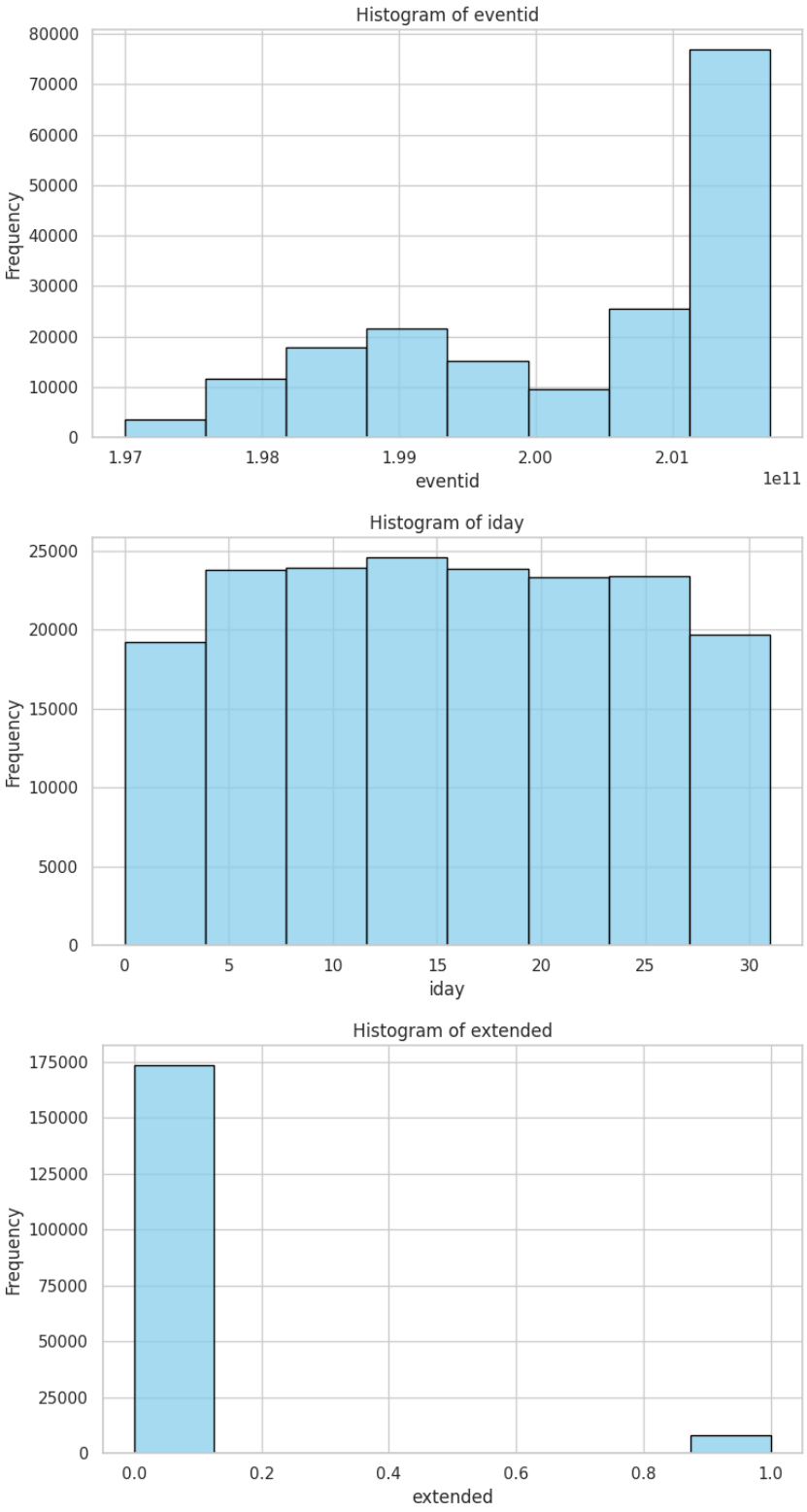
3.3 Challenges Faced

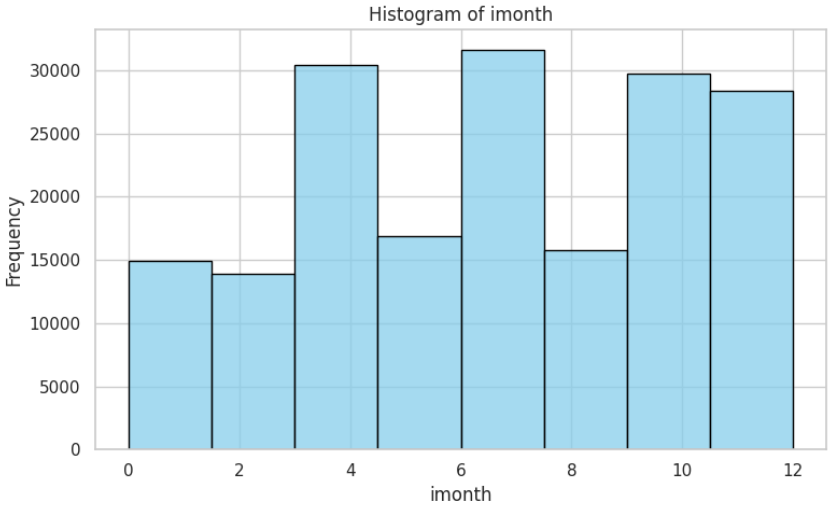
During data loading, a `DtypeWarning` appeared:

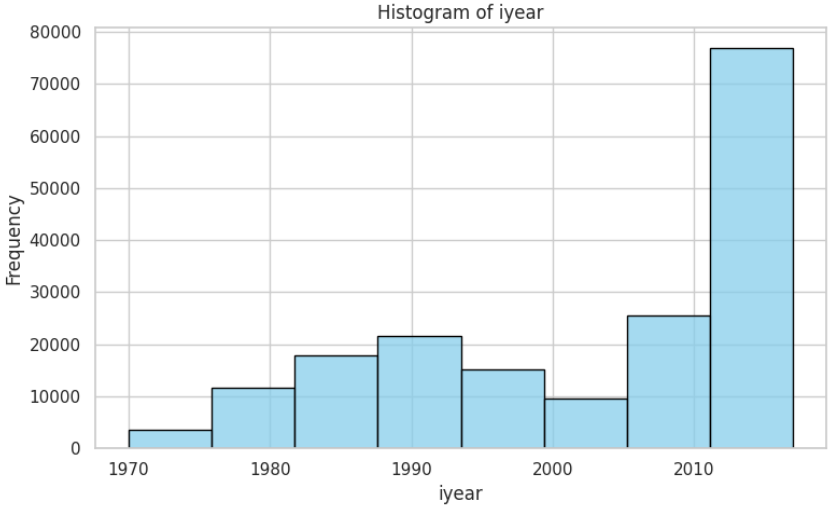
```
DtypeWarning: Columns (4,6,31,33,61,62,63,76,79,90,92,94,96,114,115,121)
have mixed types. Specify dtype option on import or set low_memory=False.
```

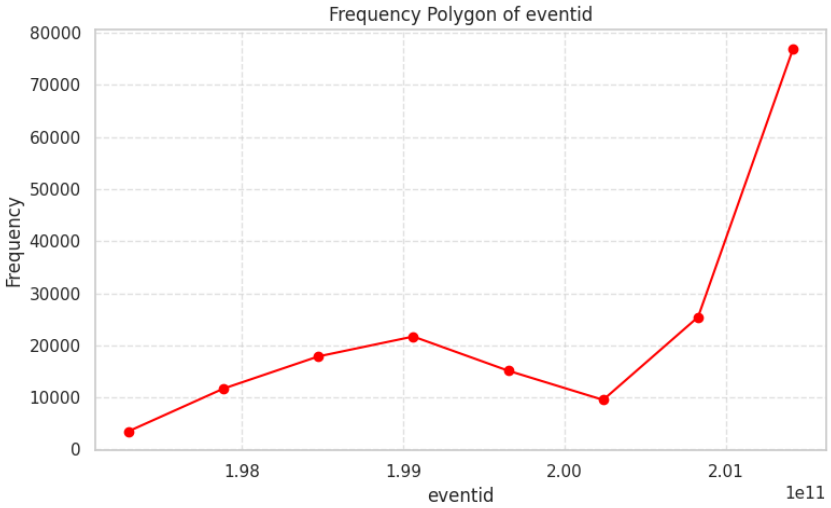
This warning indicates that some columns mix numeric and non-numeric values, forcing pandas to guess types and potentially increasing memory use. For this exploratory analysis, the data still loaded and the frequency work proceeded without issues, so I accepted the default behavior. For more rigorous or production analysis, it would be safer to specify column dtypes explicitly or set `low_memory=False` and then clean or convert types to avoid misinterpretation.

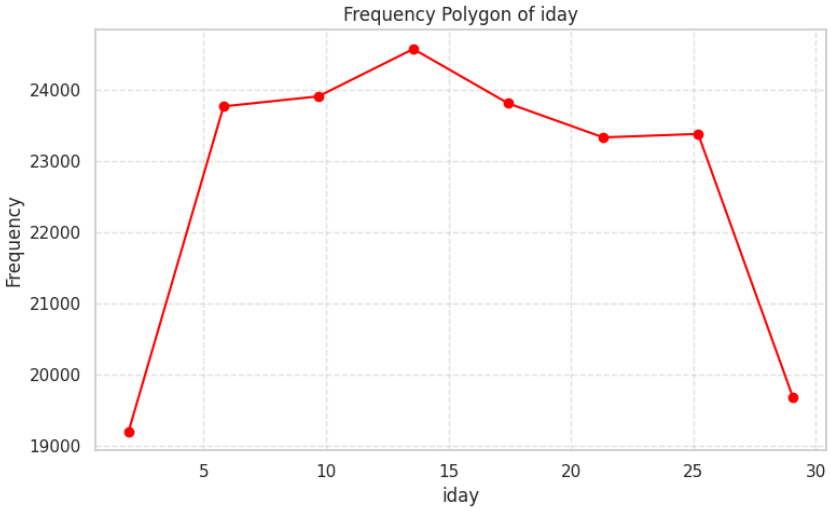
3.4 Graphical Representation

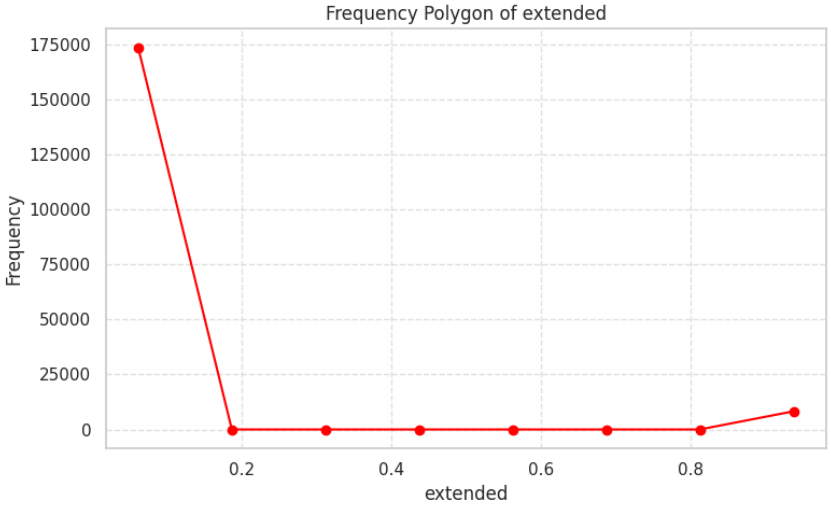


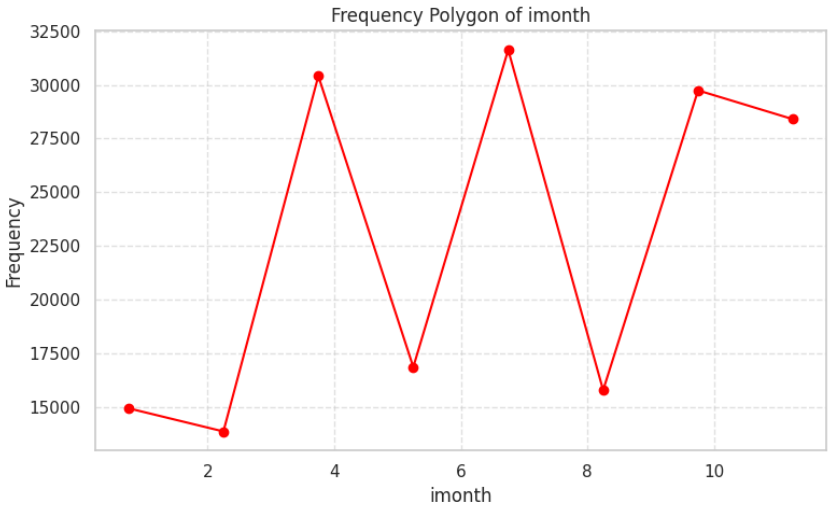

Histogram of imonth
Frequency
imonth
0
5000
10000
15000
20000
25000
30000
0
2
4
6
8
10
12

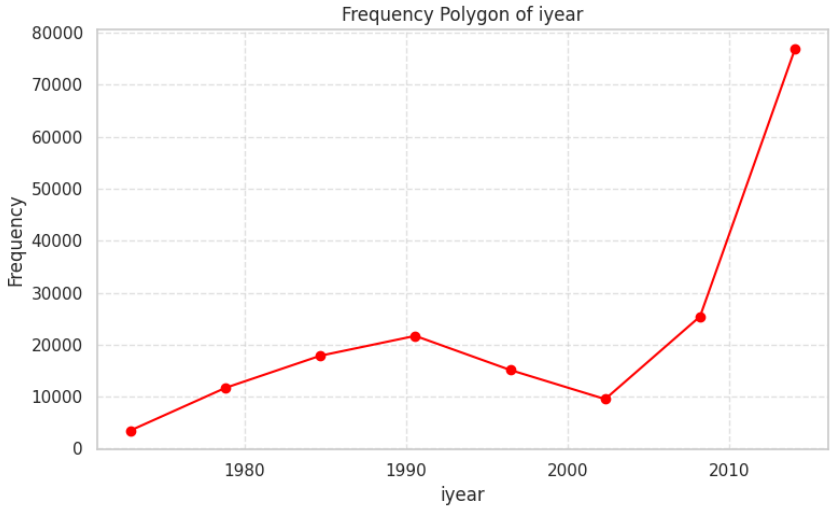

Histogram of iyear
Frequency
iyear
0
10000
20000
30000
40000
50000
60000
70000
80000
1970
1980
1990
2000
2010

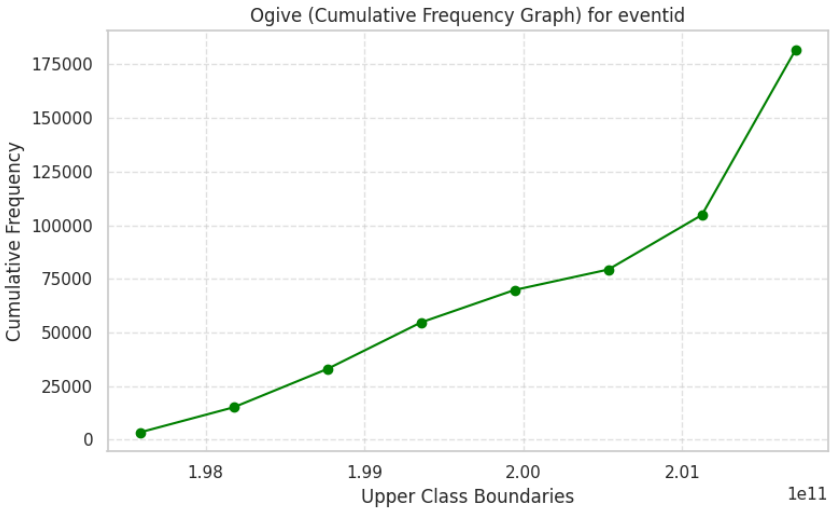

Frequency Polygon of eventid
Frequency
eventid
1e11
0
10000
20000
30000
40000
50000
60000
70000
80000
1.98
1.99
2.00
2.01

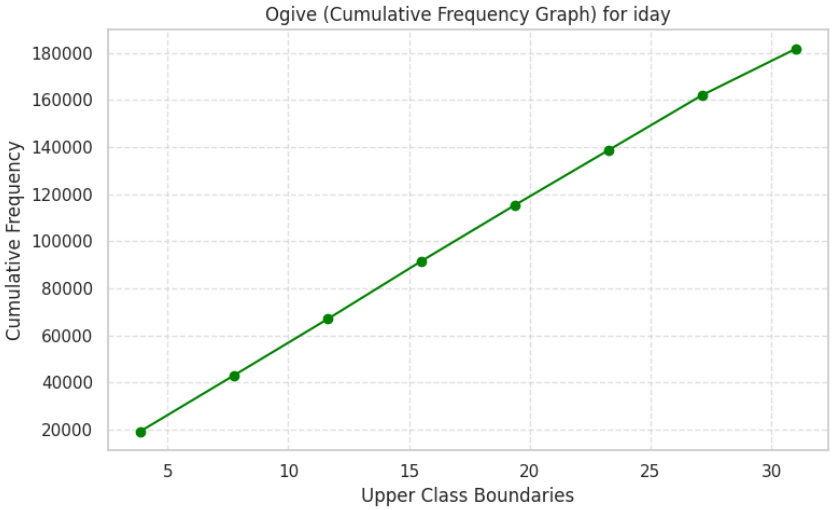

Frequency Polygon of iday
Frequency
24000
23000
22000
21000
20000
19000
5
10
15
20
25
30
iday

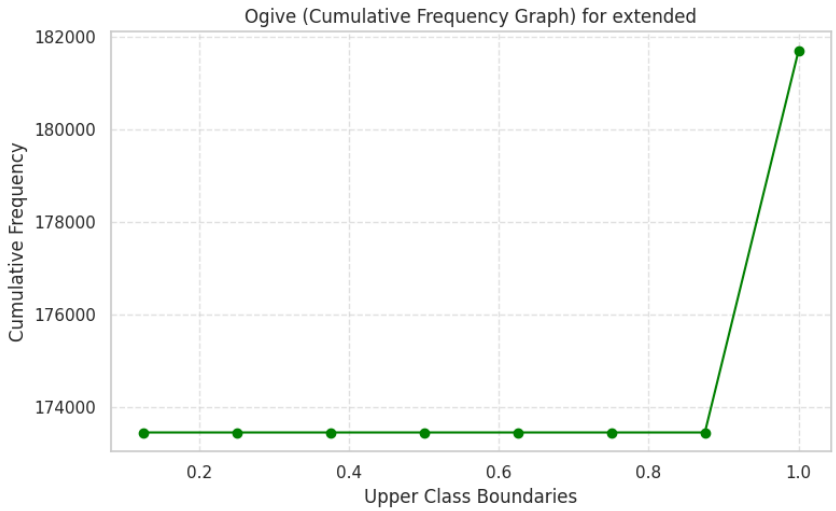

Frequency Polygon of extended
Frequency
175000
150000
125000
100000
75000
50000
25000
0
0.2
0.4
0.6
0.8
extended

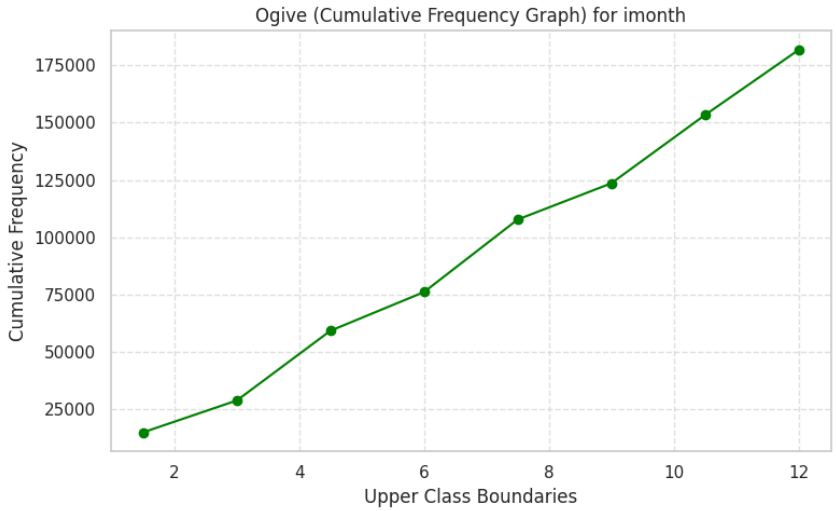

Frequency Polygon of imonth
Frequency
32500
30000
27500
25000
22500
20000
17500
15000
2
4
6
8
10
imonth

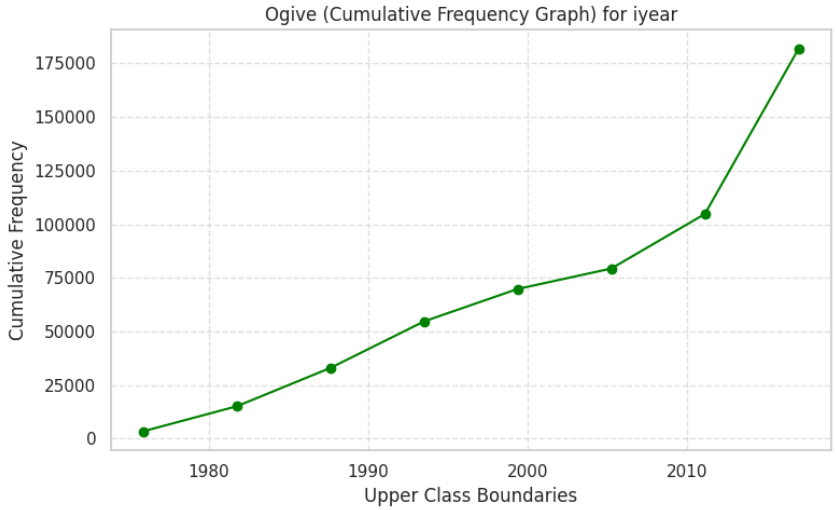

Frequency Polygon of iyear
Frequency
iyear
80000
70000
60000
50000
40000
30000
20000
10000
0
1980
1990
2000
2010


Ogive (Cumulative Frequency Graph) for eventid
Cumulative Frequency
Upper Class Boundaries
1e11
175000
150000
125000
100000
75000
50000
25000
0
1.98
1.99
2.00
2.01


Ogive (Cumulative Frequency Graph) for iday
Cumulative Frequency
Upper Class Boundaries
180000
160000
140000
120000
100000
80000
60000
40000
20000
5
10
15
20
25
30


Ogive (Cumulative Frequency Graph) for extended
Cumulative Frequency
182000
180000
178000
176000
174000
0.2
0.4
0.6
0.8
1.0
Upper Class Boundaries


Ogive (Cumulative Frequency Graph) for imonth
Cumulative Frequency
175000
150000
125000
100000
75000
50000
25000
2
4
6
8
10
12
Upper Class Boundaries


Ogive (Cumulative Frequency Graph) for iyear
Cumulative Frequency
175000
150000
125000
100000
75000
50000
25000
0
1980
1990
2000
2010
Upper Class Boundaries

4 Milestone 4 – Measures of Central Tendency and Dispersion

4.1 Column Selection Justification

The numerical columns were selected to focus on meaningful quantitative measures while avoiding identifiers, redundant fields, and low-information codes.

1. **Exclusion of identifier/temporal columns**: `eventid`, `iyear`, `imonth`, `iday`, and `approxdate` were excluded because they act as IDs or time markers rather than direct measures of incident characteristics. In contrast, `latitude` and `longitude` were kept as true numeric coordinates.

2. **Exclusion of low-cardinality numeric codes**: Columns such as `extended`, `vicinity`, `crit1`, `crit2`, `crit3`, and similar flags were dropped from this stage when they had very few unique values (threshold `nunique() <= 15`), since they behave more like categorical indicators than continuous variables.

3. **Inclusion of quantitative measures**: Variables such as `nkill`, `nwound`, `propvalue`, `ransomamt`, `nperps`, `nhours`, and `ndays` were kept because they quantify impact, duration, perpetrators, or economic loss.

4. **Handling missing values**: Columns with many missing entries (e.g., `targtype2`, `natlty2`, `targtype3`, `natlty3`) were retained, with missing values imputed using the column mean. This preserves as much numeric information as possible while enabling calculation of descriptive statistics on the full dataset.

5. **Exclusion of redundant "US-specific" columns**: Fields such as `nkillus`, `nwoundus`, and `ransomamtus` were excluded because their general counterparts (`nkill`, `nwound`, `ransomamt`) were already included, and the US-specific versions would be subsets.

Overall, this selection strategy aimed to keep numeric columns that genuinely describe the scale and impact of attacks while limiting redundancy and purely coded flags that would distort continuous summaries.

4.2 Analysis and Conclusion

Below are the main descriptive patterns for several key numerical columns and the impact of mean imputation.

4.2.1 `country` (Country Code)

Central tendency: Mean = 131.97, Median = 98.0, Mode = 95.0. The mean is higher than both median and mode, indicating a right-skewed distribution: more incidents appear in lower-numbered country codes, but a few high codes pull the mean to the right.

Dispersion and shape: Standard deviation = 112.41, variance = 12637.03, showing substantial spread, as expected in a global dataset spanning many countries. The histogram would peak near 95 with a long right tail. There were no missing values, so imputation did not affect this column.

4.2.2 `latitude`

Central tendency: Mean = 23.50, Median = 31.13, Mode = 33.30. The mean is lower than the median and mode, suggesting left-skewness: many incidents occur at higher (more northern) latitudes, while some lower or negative latitudes pull the mean downward.

Dispersion and shape: Standard deviation = 18.33, variance = 336.17, indicating a moderate spread across latitude bands. The histogram would peak at positive mid-latitudes with a tail toward lower values. Missing values were imputed with the mean, which slightly centralizes the distribution and may reduce the apparent variance and skew compared with the raw data.

4.2.3 `longitude`

Central tendency: Mean = -458.70, Median = 43.14, Mode = 44.37. The huge gap between the mean and the median/mode points to severe skewness or data errors: longitudes normally lie between -180 and 180, so a mean near -458.70 is not plausible.

Dispersion and shape: Standard deviation ≈ 202194.6, variance $\approx 4.09 \times 10^{10}$, far beyond what genuine longitude values should produce. This suggests extreme outliers or incorrect entries that dominate the mean and spread. In a histogram, most values likely cluster in the valid longitude range, with a few extremely negative points creating an enormous left tail. Mean imputation would propagate this bias, as missing values are replaced by an already distorted mean.

4.2.4 `nkill` (Number Killed)

Central tendency: Mean = 2.40, Median = 1.00, Mode = 0.00. The mean being above both median and mode shows a strongly right-skewed distribution: most attacks cause few or no deaths, while a small set of high-fatality incidents lifts the mean.

Dispersion and shape: Standard deviation = 11.21, variance = 125.74, indicating substantial spread relative to the mean. The histogram would be heavily concentrated at

0 and low counts, with a long right tail for severe events. Missing values were imputed with the mean (2.40), which introduces non-integer values and may slightly smooth the spike at zero by adding mid-level values to previously missing entries.

4.2.5 `nwound` (Number Wounded)

Central tendency: Mean = 3.17, Median = 0.00, Mode = 0.00. As with `nkill`, many incidents produce no injuries, but some produce many, creating a strong right-skew.

Dispersion and shape: Standard deviation = 34.30, variance = 1176.34, show very high variability, reflecting the wide range of attack severities. The histogram would mirror `nkill` with a dominant bar at zero and a long tail. Mean imputation at 3.17 fills missing values with a moderate positive number, which can slightly lift the lower end and further blur the distinction between zero and low counts.

4.2.6 `propvalue` (Property Value)

Central tendency: Mean = 208811.87, Median = 208811.87, Mode = 208811.87. Having all three measures equal signals that the imputed mean value appears very frequently—most likely because many missing entries were replaced by this mean, making it both the center and the most common value.

Dispersion and shape: Standard deviation $\approx 7.19 \times 10^6$, variance $\approx 5.17 \times 10^{13}$, indicate extreme spread, suggesting either very large true losses for some events or influential outliers. The histogram would likely show a huge spike at the imputed mean with relatively fewer original values scattered elsewhere. This pattern underlines how heavy mean imputation can distort the perceived distribution of property damage.

4.2.7 Other Columns and Overall Characteristics

Several other variables, especially `targtype`, `targsubtype`, `natlty`, `weapsubtype`, `nhours`, `ndays`, `ransomamt`, and `nreleased`, show substantial gaps between mean, median, and mode and often very high variance. In some cases, central tendencies become negative, which is not meaningful for durations or counts and likely reflects coded values such as -9 or -99 for "unknown" being treated as numeric.

Global patterns:

- Many important variables (e.g., `nkill`, `nwound`) are strongly right-skewed with lots of zeros and a few large outliers.
- Variability is high in several columns (`longitude`, `propvalue`, `nkill`, `nwound`), emphasizing the heterogeneous nature of global terrorism incidents.

- Mean imputation has a major impact where missingness is heavy, especially in `propvalue` and coded fields like `nhours`, `ndays`, `ransomamt`, and `nreleased`, sometimes producing misleading central values.
- Data quality issues are evident in variables such as `longitude` and in negative "duration" or "count" values, which likely combine real measurements with placeholder codes.

Overall, the dataset is numerically rich but strongly skewed and affected by outliers and coding conventions, which calls for robust methods and careful preprocessing in any further analysis.

4.3 Challenges Faced

Several practical challenges arose while computing and interpreting descriptive statistics.

4.3.1 High Dimensionality and Missing Data

With 135 columns and many missing values, selecting truly numerical variables and deciding how to handle missingness was non-trivial. Columns such as `targtype2`, `natlty2`, and `weapsubtype2` are often missing by design when there is no second or third target. Mean imputation enabled full-sample summaries but may introduce bias when missingness is extensive or systematic.

4.3.2 Interpreting Numeric Codes

Many fields (`country`, `targtype1`, `weapsubtype1`) are numeric codes for categories. Treating them as continuous variables and interpreting their means or variances can be misleading. They were included here for completeness, but they are better analyzed via frequency tables and modes, with a categorical perspective.

4.3.3 Skewed Distributions and Outliers

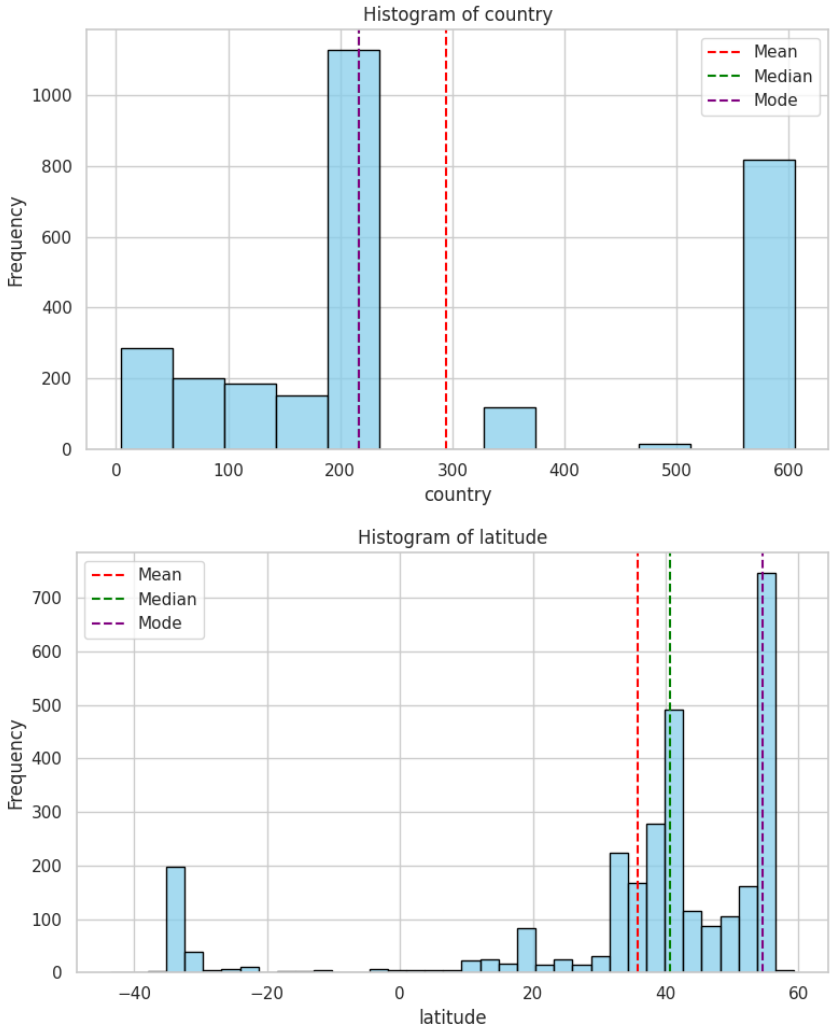
Variables like `nkill`, `nwound`, `propvalue`, and `ransomamt` are highly skewed, with many zeros and a few extreme values. In these cases, the mean is far less robust than the median, and large gaps between mean, median, and mode are expected. Visualizing histograms and noting this divergence was essential for understanding the true shape of the data.

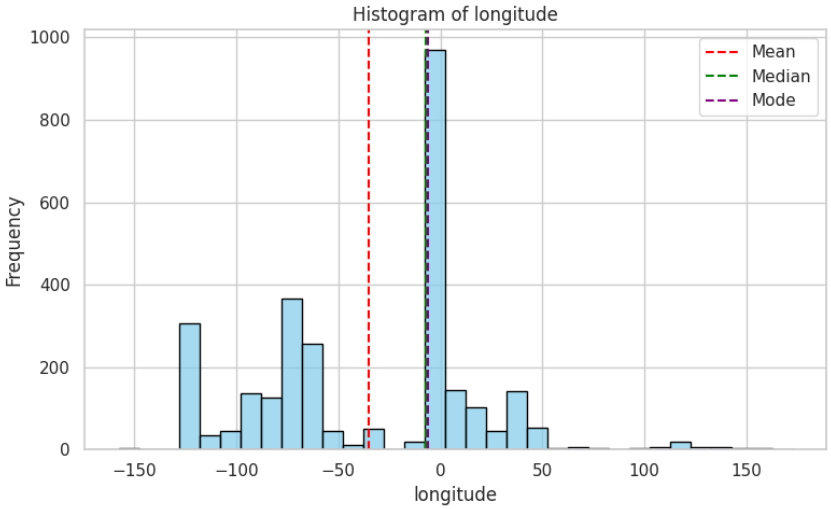
4.3.4 Negative Placeholder Values

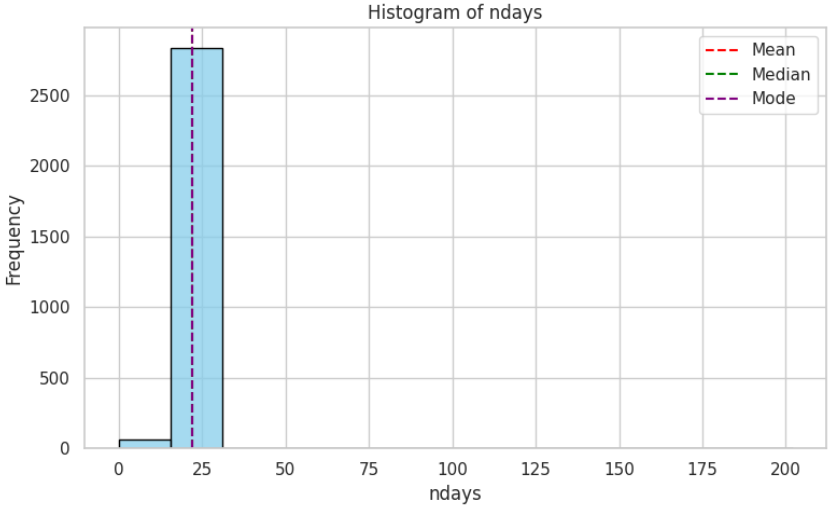
Columns such as `nperps`, `nhours`, `ndays`, and `nreleased` contain negative values that likely represent "unknown" or "not applicable" codes (e.g., -9 or -99). Treating these as

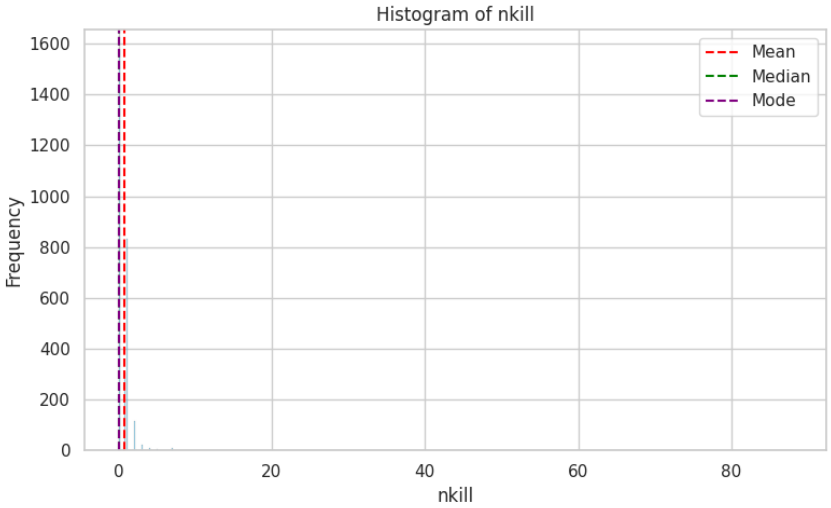
genuine numbers and then imputing with the mean can produce central tendencies that are not interpretable in real-world terms (for example, a mean of -65 for `nperps`). A more robust workflow would recode these placeholders to `NaN` before analysis or exclude them from certain calculations.

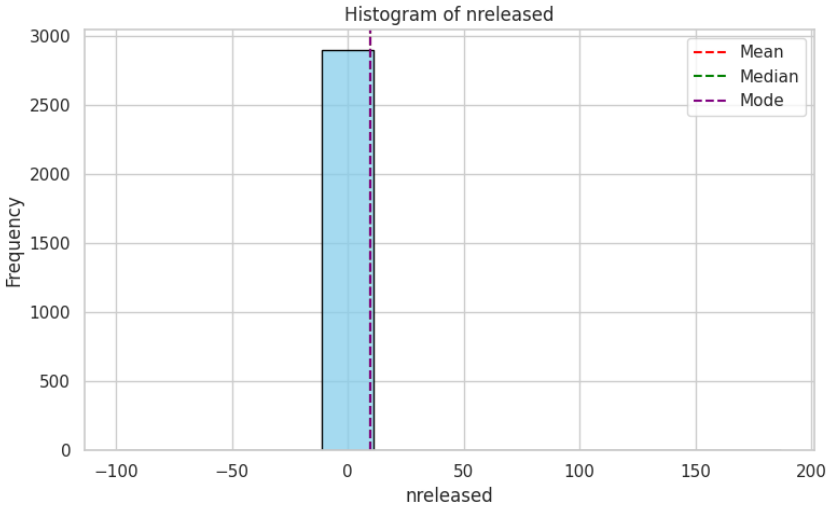
4.4 Graphical Representation

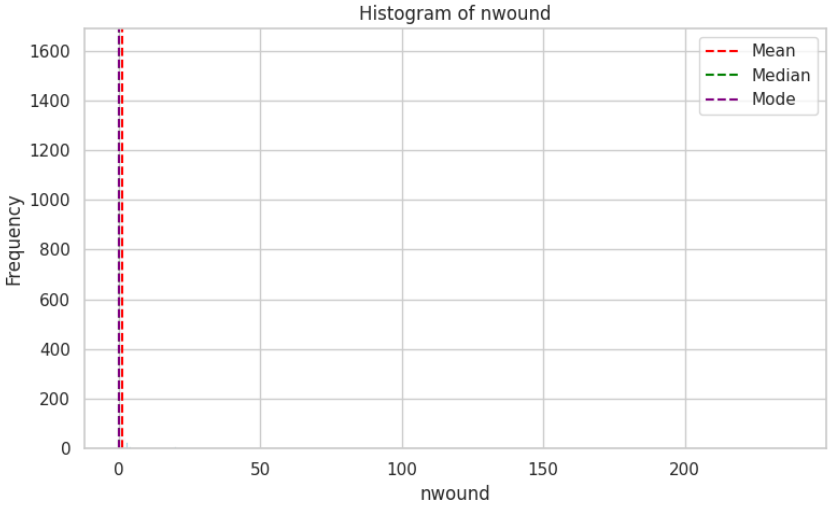


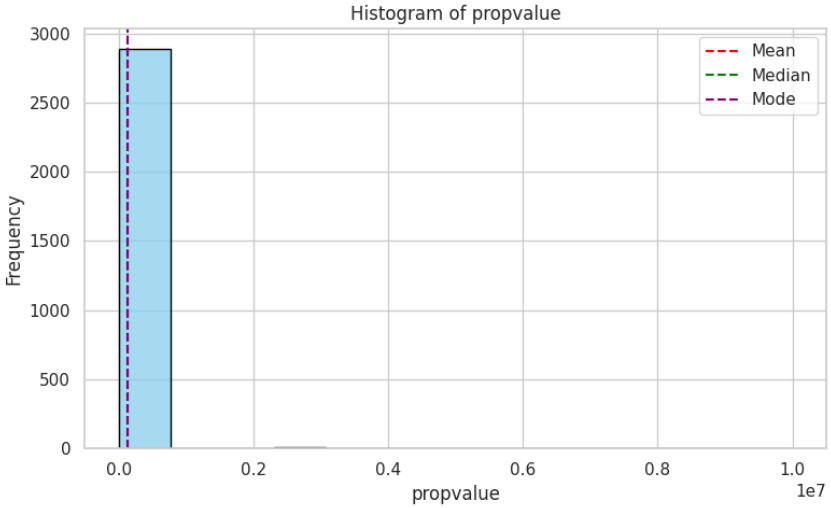

Histogram of longitude
Mean
Median
Mode
Frequency
longitude
1000
800
600
400
200
0
−150
−100
−50
0
50
100
150

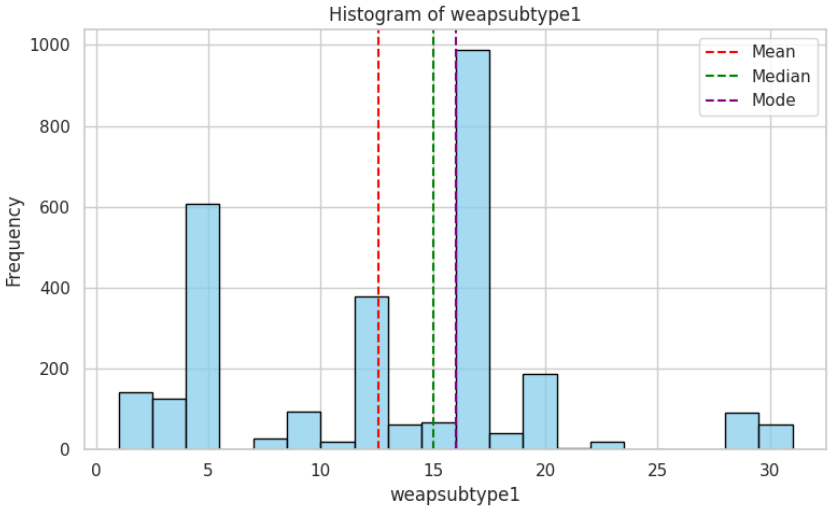

Histogram of ndays
Mean
Median
Mode
Frequency
ndays
2500
2000
1500
1000
500
0
0
25
50
75
100
125
150
175
200

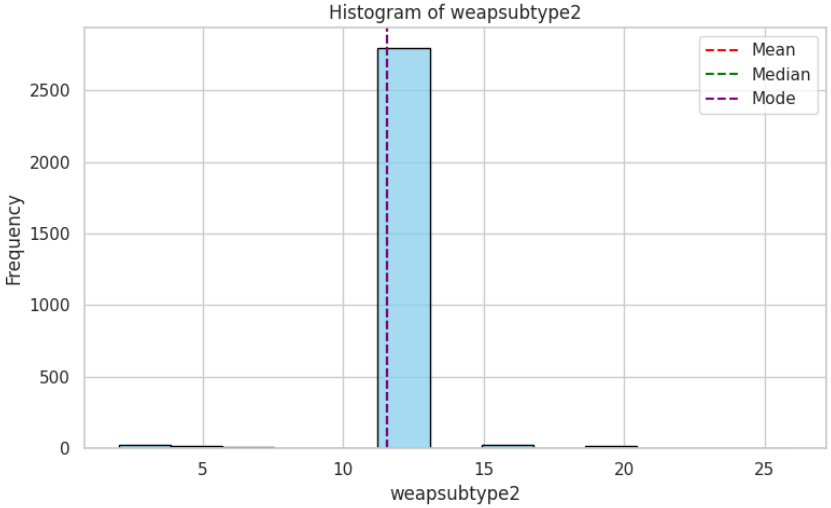

Histogram of nkill
Mean
Median
Mode
Frequency
nkill
1600
1400
1200
1000
800
600
400
200
0
0
20
40
60
80

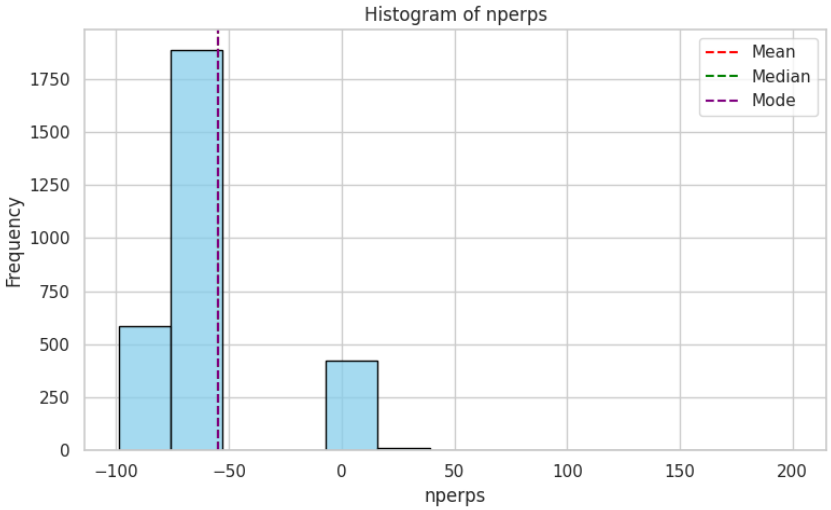

Histogram of nreleased
Mean
Median
Mode
Frequency
nreleased
3000
2500
2000
1500
1000
500
0
−100
−50
0
50
100
150
200

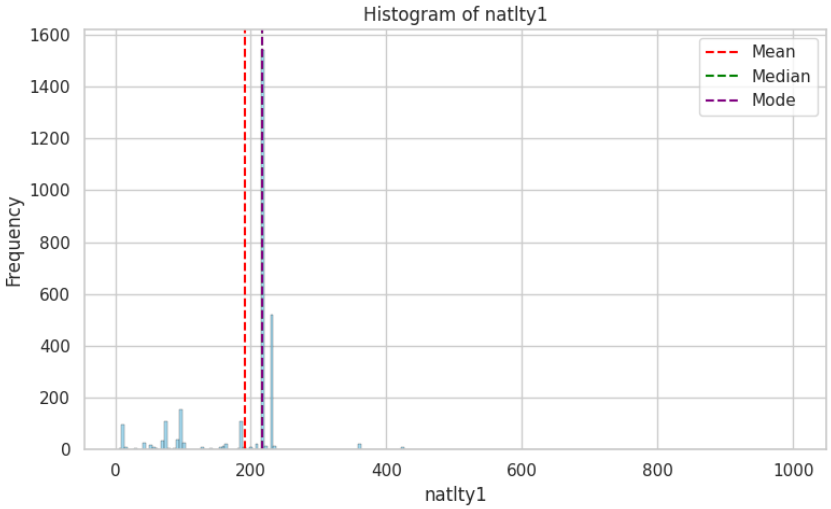

Histogram of nwound
Mean
Median
Mode
Frequency
nwound
1600
1400
1200
1000
800
600
400
200
0
0
50
100
150
200

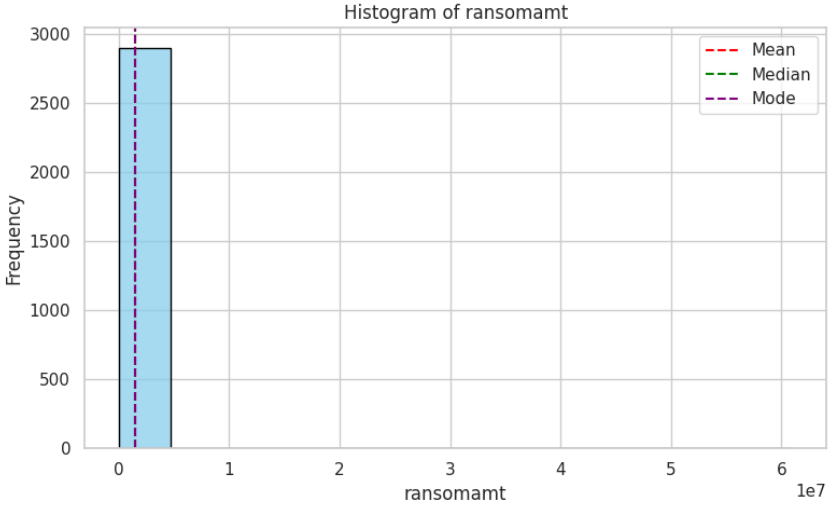

Histogram of propvalue
Mean
Median
Mode
Frequency
propvalue
1e7
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0

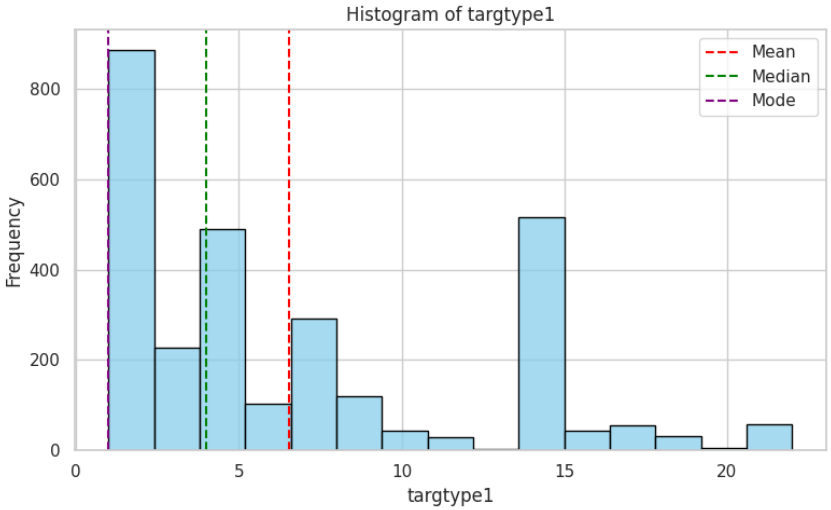

Histogram of weapsubtype1
Mean
Median
Mode
Frequency
weapsubtype1
0
200
400
600
800
1000
0
5
10
15
20
25
30

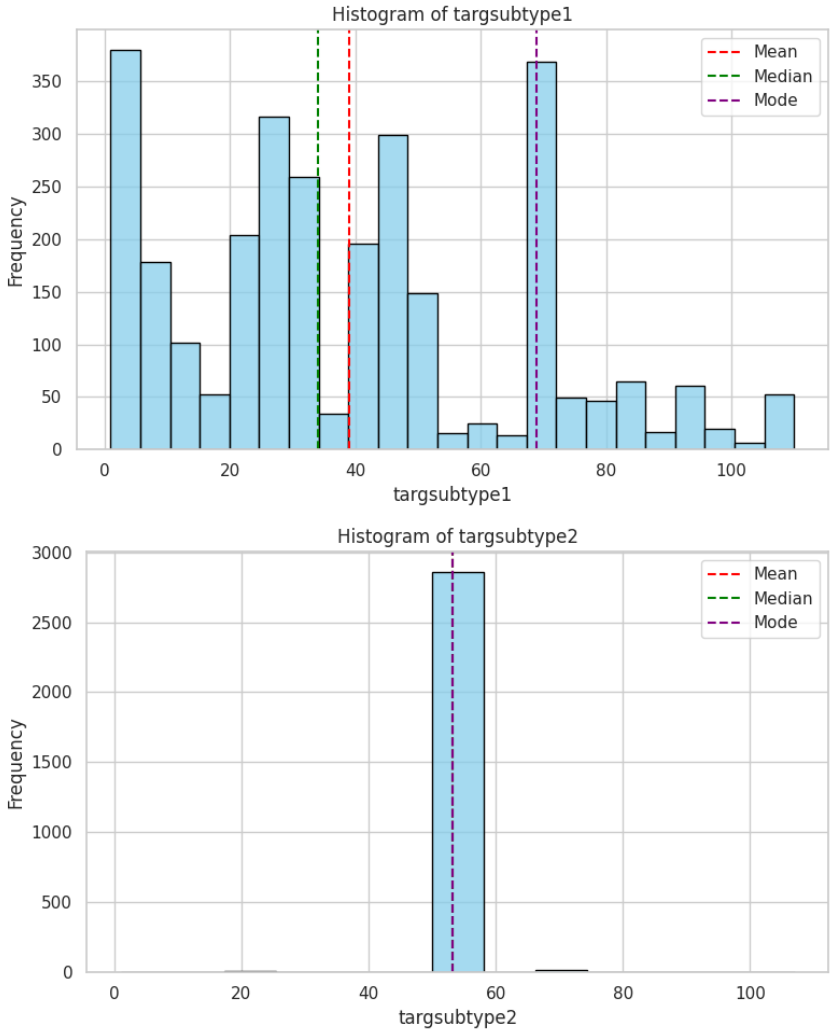

Histogram of weapsubtype2
Mean
Median
Mode
Frequency
weapsubtype2
0
500
1000
1500
2000
2500
5
10
15
20
25


Histogram of nperps
Mean
Median
Mode
Frequency
nperps
0
250
500
750
1000
1250
1500
1750
−100
−50
0
50
100
150
200


Histogram of natlty1
Mean
Median
Mode
Frequency
natlty1
0
200
400
600
800
1000
1200
1400
1600


Histogram of ransomamt
Mean
Median
Mode
Frequency
ransomamt
1e7
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6


Histogram of targtype1
Mean
Median
Mode
Frequency
targtype1
0
200
400
600
800
5
10
15
20



5 Milestone 5 – Introduction to Probability

5.1 Column Selection Justification and Event Defining Justification

5.1.1 Column Selection

I selected the following columns to capture human impact, operational outcome, and material damage:

- `nkill`: Number of fatalities, central for defining lethality-based events.
- `nwound`: Number of wounded, complementing `nkill` to describe total human cost.
- `success`: Binary indicator (1/0) of operational success, used to study how often attacks achieve their immediate goals.
- `property`: Indicates property damage (1/0 or positive value), capturing material and economic impact.

- `extended`: Binary flag for incidents that last beyond a single day, representing prolonged or multi-phase operations.

Together, these columns allow probability-based analysis of severity, casualties, success, damage, and duration.

5.1.2 Event Definitions

I grouped events into five sets to study different aspects of attacks.

Event Set 1: Fatality-Based Severity

- **Fatal Attack** (A_1): `nkill > 0`. Any attack with at least one death.
- **Multiple Fatalities** (B_1): `nkill >= 5`. Attacks with five or more deaths, representing highly lethal incidents.
- **No Fatalities** (C_1): `nkill == 0`. Incidents with no recorded deaths.

Event Set 2: Human Impact vs Operational Outcome

- **Any Casualty** (A_2): `(nkill.fillna(0) + nwound.fillna(0)) > 0`. At least one killed or wounded; missing values treated as 0 for summation.
- **Successful Attack** (B_2): `success == 1`. Operationally successful incidents.
- **Property Damage** (C_2): `property > 0`. Attacks causing damage to property.

Event Set 3: Injury-Focused Events

- **Any Injury** (A_3): `nwound > 0`. At least one reported injury.
- **High Injury** (B_3): `nwound >= median_injury`. Injuries at or above the median among injury cases (median computed using `nwound > 0` only).
- **No Injury** (C_3): `nwound == 0`. Incidents with no injuries.

Event Set 4: Severity and Duration Interaction

- **Extended Attack** (A_4): `extended == 1`. Multi-day or prolonged attacks.
- **Fatal Attack** (B_4): `nkill > 0`. Same as A_1 for comparison.
- **Fatal + Injury** (C_4): `(nkill > 0) & (nwound > 0)`. Events with both deaths and injuries.

Event Set 5: Damage-Oriented Events

- **Property Damage** (A_5): `property > 0`. Same as C_2.
- **Fatal Attack** (B_5): `nkill > 0`. Same as A_1.
- **Injury Attack** (C_5): `nwound > 0`. Same as A_3.

These event sets provide a structured way to compute probabilities, intersections, and unions for different dimensions of attack impact.

5.2 Reflection and Summary

5.2.1 Key Probabilities for Each Event Set

Event Set 1: Fatality-Based Severity

- $P(A_1)$ (Fatal Attack, `nkill > 0`) = 0.458: Around 45.8% of attacks cause at least one death.
- $P(B_1)$ (Multiple Fatalities, `nkill >= 5`) = 0.109: About 10.9% have five or more deaths.
- $P(C_1)$ (No Fatalities, `nkill == 0`) = 0.485: Roughly 48.5% have no fatalities.

Multiple-fatality events are a subset of fatal attacks, and nearly half of all incidents produce no deaths, showing a wide range of lethality.

Event Set 2: Human Impact vs Operational Outcome

- $P(A_2)$ (Any Casualty) = 0.593: About 59.3% of attacks cause at least one death or injury.
- $P(B_2)$ (Successful Attack) = 0.890: Roughly 89.0% are recorded as operationally successful.
- $P(C_2)$ (Property Damage) = 0.515: Around 51.5% involve property damage.

The joint probability $P(A_2 \cap B_2) = 0.561$, while $P(A_2)P(B_2) = 0.593 \times 0.890 \approx 0.528$. Since $0.561 > 0.528$, successful attacks are more likely than expected to involve casualties, indicating a positive association between success and human impact.

Event Set 3: Injury-Focused Events

- $P(A_3)$ (Any Injury, `nwound > 0`) = 0.342: About 34.2% of attacks cause at least one injury.
- $P(B_3)$ (High Injury, `nwound >= 3`) = 0.197: Around 19.7% reach or exceed the median non-zero injury count (3).
- $P(C_3)$ (No Injury, `nwound == 0`) = 0.568: About 56.8% have no reported injuries.

High-injury events are a subset of injury events, mirroring the subset relationship seen between multiple-fatality and fatal attacks.

Event Set 4: Severity and Duration Interaction

- $P(A_4)$ (Extended Attack, `extended == 1`) = 0.045: Only about 4.5% of attacks are extended.
- $P(B_4)$ (Fatal Attack) = 0.458: Same as $P(A_1)$.
- $P(C_4)$ (Fatal + Injury) = 0.207: About 20.7% result in both fatalities and injuries.

Here $P(A_4 \cap B_4) = 0.015$, while $P(A_4)P(B_4) = 0.045 \times 0.458 \approx 0.021$. Since $0.015 < 0.021$, extended attacks are slightly less likely to be fatal than they would be under independence, hinting that prolonged incidents may often emphasize objectives other than immediate high fatality counts.

Event Set 5: Damage-Oriented Events

- $P(A_5)$ (Property Damage) = 0.515.
- $P(B_5)$ (Fatal Attack) = 0.458.
- $P(C_5)$ (Injury Attack) = 0.342.

For property damage and fatalities, $P(A_5 \cap B_5) = 0.180$, while $P(A_5)P(B_5) = 0.515 \times 0.458 \approx 0.236$. Since $0.180 < 0.236$, attacks involving property damage are less likely to be fatal than if these two characteristics were independent, suggesting different emphases between property-focused and highly lethal operations.

5.2.2 Cross-Set Trends and Implications

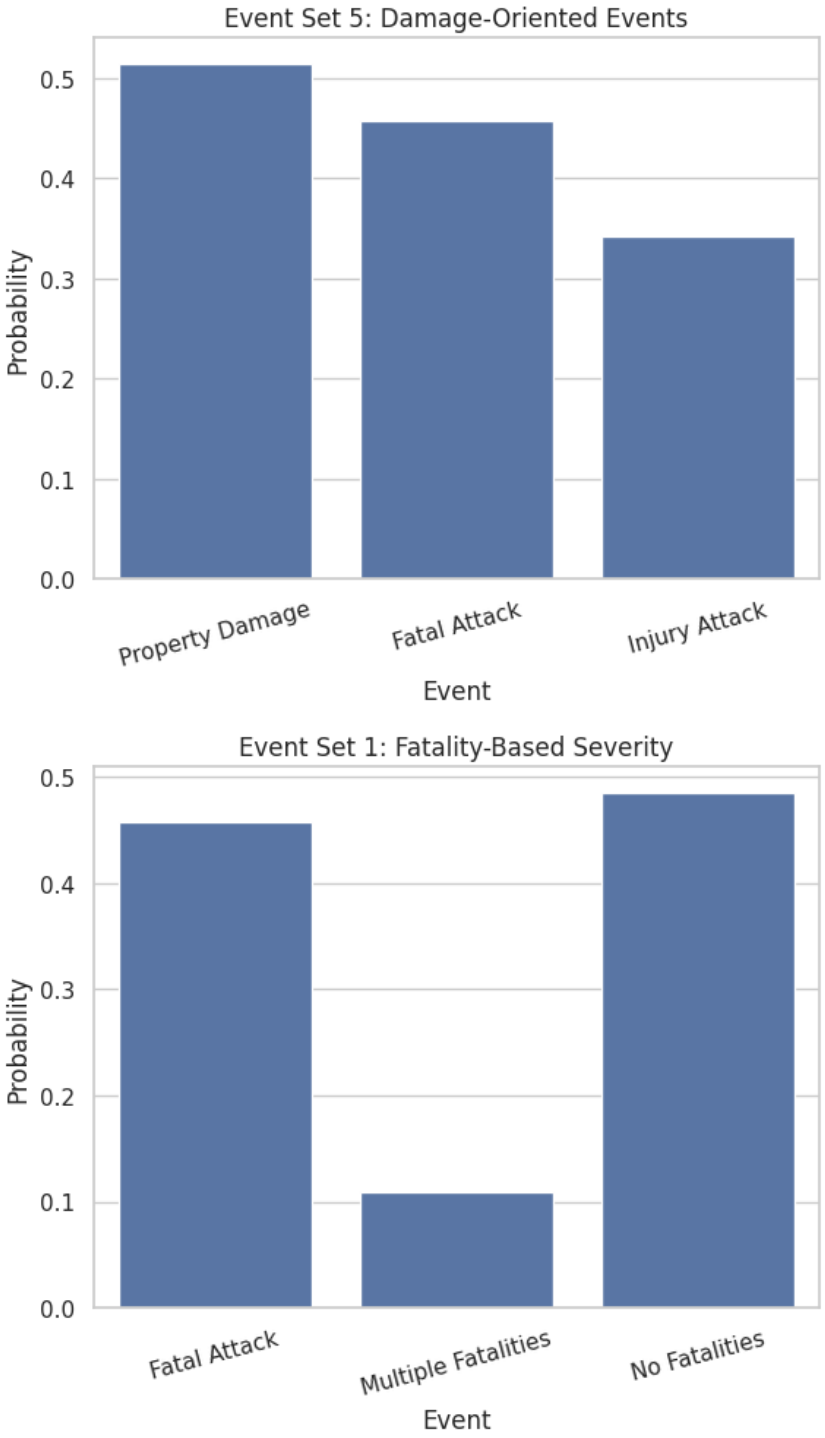
Across sets, $P(\text{Fatal Attack}) \approx 0.458$ appears consistently, while $P(\text{Any Casualty}) \approx 0.593$ shows that injuries broaden the pool of impactful events beyond fatalities alone. $P(\text{No Fatalities}) \approx 0.485$ and $P(\text{No Injury}) \approx 0.568$ confirm that many attacks cause no direct physical harm, highlighting the diversity of terrorist objectives.

The very high $P(\text{Successful Attack}) = 0.890$ compared with lower casualty probabilities suggests that "success" is often defined by goals such as disruption, demonstration of capability, or symbolic impact, not only by high casualty counts. Property damage, with $P(\text{Property Damage}) = 0.515$, is nearly as common as any casualty, underlining the importance of economic and infrastructural targets.

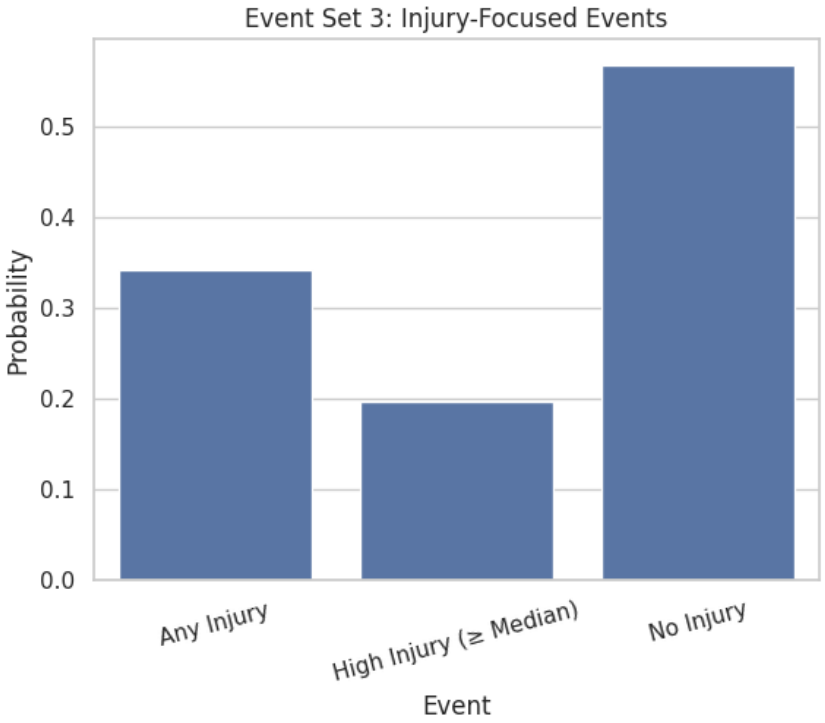
Positive association between success and casualties, along with negative associations between extended attacks and fatalities, and between property damage and fatalities, shows that different attack profiles exist: some aimed at causing casualties, others at disruption or prolonged pressure.

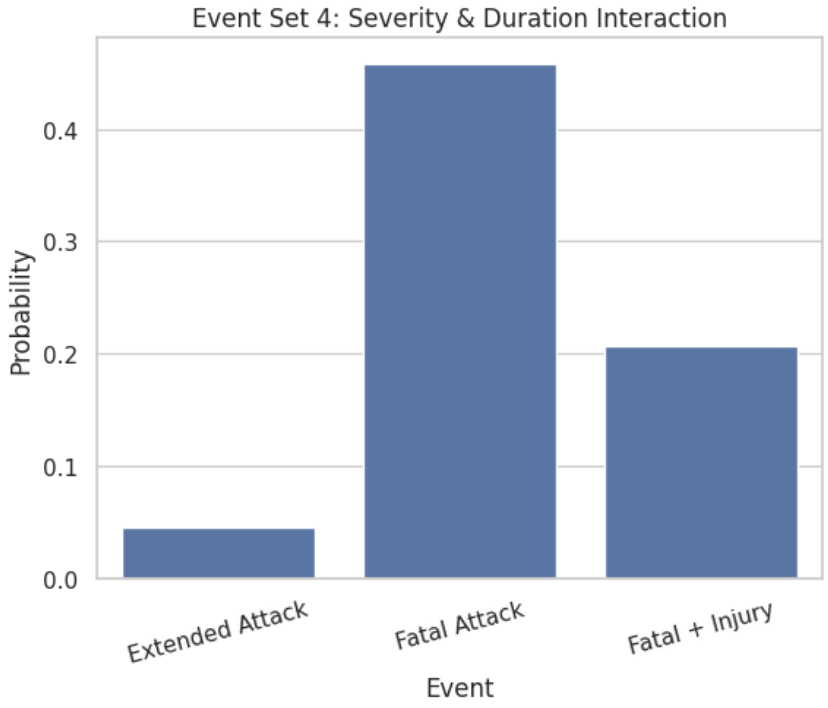
Overall, these probability-based event definitions and results provide a structured view of how often different types of harm and outcomes occur together and separately, giving a clearer picture of the varied nature of terrorist activity.

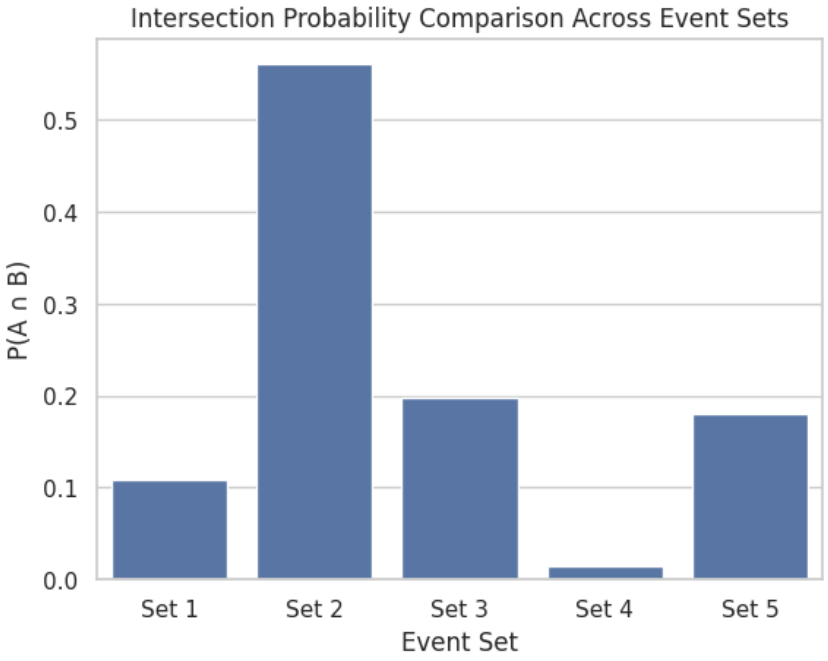
5.3 Graphical Representation




Event Set 2: Human Impact vs Operational Outcome
Probability
0.0
0.2
0.4
0.6
0.8
Any Casualty
Successful Attack
Property Damage
Event


Event Set 3: Injury-Focused Events
Probability
0.0
0.1
0.2
0.3
0.4
0.5
Any Injury
High Injury (≥ Median)
No Injury
Event


Event Set 4: Severity & Duration Interaction
Probability
0.0
0.1
0.2
0.3
0.4
Extended Attack
Fatal Attack
Fatal + Injury
Event


Intersection Probability Comparison Across Event Sets
P(A ∩ B)
0.0
0.1
0.2
0.3
0.4
0.5
Set 1
Set 2
Set 3
Set 4
Set 5
Event Set

6 Milestone 6 – Conditional Probability, Independence Check, Bayes' Rule and Normal Distributions

6.1 Column Selection and Event Definition Justification

6.1.1 Column Selection Justification

I chose a mix of categorical and numerical variables that describe attack tactics, targets, geography, and human impact.

Categorical variables:

- `attacktype1_txt`: Primary attack type (e.g., Bombing/Explosion, Armed Assault). Captures the tactic used.
- `weaptype1_txt`: Primary weapon type. Helps assess methods and potential lethality.
- `targtype1_txt`: Primary target category (e.g., business, government, civilians), revealing strategic choices and intent.
- `region_txt`: Geographic region of the attack, used to compare regional patterns and hotspots.
- `success`: Binary indicator of whether the attack was successful (1) or not (0), used to study operational outcomes.

Numerical variables:

- `nkill`: Number of fatalities, a direct measure of severity and human cost.
- `nwound`: Number of wounded, complementing `nkill` to describe overall casualties.

These variables together support conditional probability, independence checks, Bayes' rule, and normality comparisons focused on severity and success.

6.1.2 Event Definition Justification

Three core events were defined to capture different severity and outcome thresholds. [

- **Event A: Fatal Attack** (A)
  Definition: `nkill > 0`.
  Rationale: Distinguishes any attack with at least one death from non-fatal incidents, marking a key boundary in human impact.

- **Event B: Successful Attack** (B)
  Definition: `success == 1`.
  Rationale: Directly uses the dataset's success flag to represent whether the attack's immediate objectives were achieved, providing a standardized measure of operational effectiveness.

- **Event C: Mass-Casualty Event** (C)
  Definition: `nkill >= 5.0`.
  Rationale: Exploratory analysis showed that, among fatal attacks, the 90th percentile of `nkill` is 5.0. Using `nkill >= 5.0` marks the top 10% most lethal attacks as mass-casualty events within this dataset's distribution.

6.2 Analysis and Reflection

6.2.1 Categorical Variable Frequencies

Frequency tables show clear dominance of certain categories:

- `attacktype1_txt`: *Bombing/Explosion* is by far the most common attack type, followed by *Armed Assault* and *Assassination*, indicating a strong reliance on explosive and armed methods.

- `weaptype1_txt`: *Explosives* are the most frequent weapons, with *Firearms* also heavily represented, reinforcing the destructive nature of prevalent tactics.

- `targtype1_txt`: *Private Citizens & Property* and *Military* are the most common targets, with *Police* and *Government (General)* also frequent, showing both civilians and state actors are primary focuses.

- `region_txt`: *Middle East & North Africa* and *South Asia* have the highest counts, highlighting them as major hotspots; *Western Europe* and *South America* also show notable activity.

- `success`: Most attacks are recorded as successful, with a success rate of about 89%, indicating that perpetrators often achieve their immediate objectives.

6.2.2 Numerical Variables and Histograms (`nkill`, `nwound`)

For `nkill`:

- Mean = 2.40, Median = 0.00, StdDev = 11.55, Max = 1570.

- The median of zero means most attacks cause no deaths, while a few extremely deadly events push the mean and standard deviation upwards.

For `nwound`:

- Mean = 3.29, Median = 0.00, StdDev = 35.03.
- The pattern mirrors `nkill`, with many incidents causing no injuries and a small number causing very large numbers of wounded.

Histograms for both variables confirm strong right-skewness and heavy tails: large bars at zero (or low counts) and very long tails for extreme events. This visual evidence supports the numerical summary that a minority of attacks drive much of the casualty burden.

6.2.3 Key Probabilities and Conditional Probabilities

Using the defined events:

- $P(A)$ (Fatal Attack, `nkill > 0`) = 0.4581: About 45.81% of attacks are fatal.
- $P(B)$ (Successful Attack, `success == 1`) = 0.8896: Roughly 88.96% are successful.
- $P(C)$ (Mass-Casualty, `nkill >= 5`) = 0.1090: Around 10.90% qualify as mass-casualty events.
- $P(A \mid B) = 0.4955$: Given success, the probability the attack is fatal rises to 49.55%, slightly above $P(A)$.
- $P(C \mid A) = 0.2380$: Among fatal attacks, 23.80% are mass-casualty, so about one in four fatal events reaches this high-impact level.

These values show that while less than half of all attacks are fatal, a substantial subset of fatal attacks generate very high casualty counts.

6.2.4 Independence Check: Fatal Attack (A) and Successful Attack (B)

- $P(A \cap B) = 0.4408$.
- $P(A)P(B) = 0.4581 \times 0.8896 \approx 0.4075$.

The absolute difference $|P(A \cap B) - P(A)P(B)| = 0.033$ exceeds the independence tolerance of 0.01, so A and B are not independent.

- Because $P(A \mid B) = 0.4955$ is greater than $P(A) = 0.4581$, successful attacks are more likely to be fatal than attacks overall.

This dependence suggests that operational success and lethality are linked: when attacks achieve their objectives, they more often result in deaths.

6.2.5 Bayes' Rule: $P(B \mid A)$

Bayes' rule gives:

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)}.$$

Using the observed probabilities yields $P(B \mid A) = 0.9622$, which exactly matches the empirically calculated $P(B \mid A) = 0.9622$.

This means that if an attack is fatal, there is about a 96.22% chance it is classified as successful. In other words, almost all fatal attacks are also successful operations from the perpetrators' perspective, and the Bayes' rule check confirms the internal consistency of the probability estimates.

6.2.6 Normal Distribution Analysis for `nkill`

A notional normal distribution with mean $\mu = 2.40$ and standard deviation $\sigma = 11.55$ was used as a reference for `nkill`.

- Benchmarks like $P(X > \mu) = 0.5000$ and $P(\mu - \sigma < X < \mu + \sigma) \approx 0.6827$ come from the standard normal distribution.
- In practice, `nkill` does not match these benchmarks well because of its heavy right-skew and concentration at zero. Many data points fall below $\mu - 2\sigma$, which is not meaningful here since fatalities cannot be negative.

The "normality comment" summarizes this mismatch: fatalities are right-skewed with heavy tails, and the mean (2.40) sits far above the median (0.00). Normal curves can serve as rough theoretical guides in some risk modeling contexts, but they are clearly imperfect for this casualty distribution, especially in the extremes.

6.2.7 Overall Summary and Implications

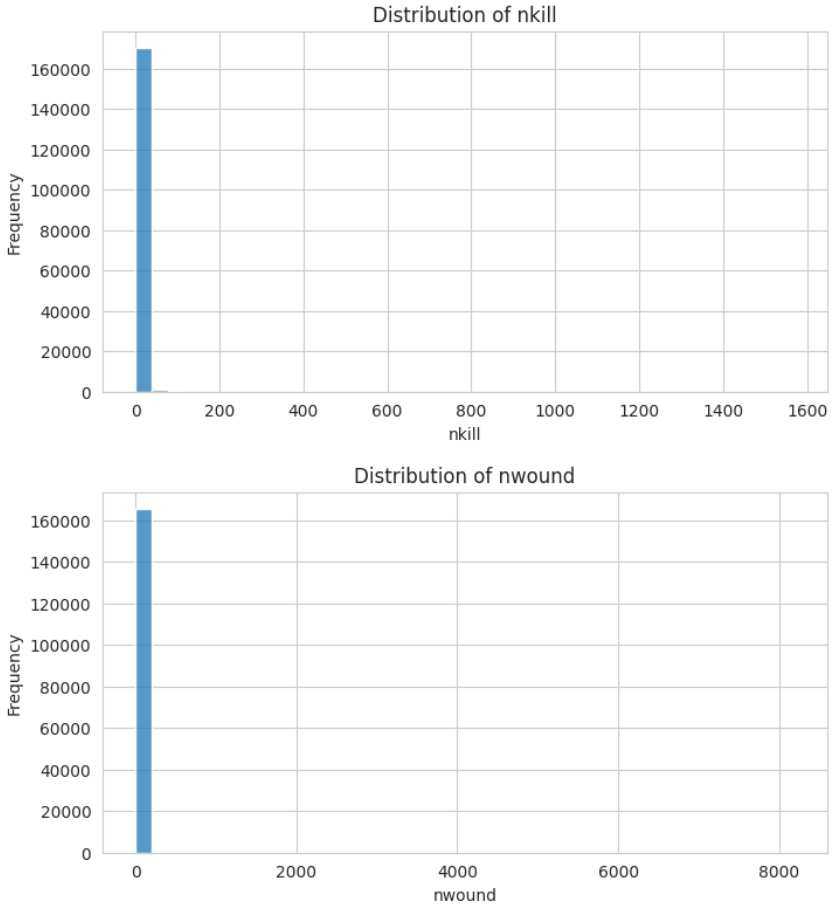
The analysis of these events and probabilities leads to several key insights:

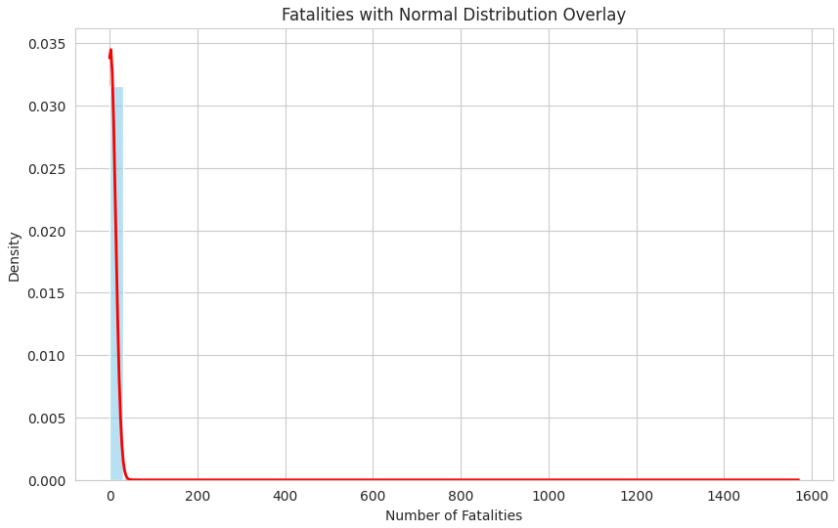
- **High success, mixed lethality**: Attacks are often successful in operational terms (nearly 89%), but less than half are fatal and many cause no casualties, showing a wide spectrum of outcomes.
- **Impact of extreme events**: Most attacks have few or no casualties, yet a relatively small group of mass-casualty incidents drives much of the aggregate fatality and injury totals.
- **Dependence between success and fatalities**: Successful attacks are more likely to be fatal, and fatal attacks are almost always successful, revealing a strong connection between operational outcome and human impact.

- **Limits of normal approximations**: The severe skew and heavy tails in `nkill` and `nwound` make simple normal-based rules unreliable, especially for tail risk. More robust or specialized models would be needed for accurate casualty risk assessment.

These findings support more nuanced counter-terrorism planning that considers not only whether attacks occur and succeed, but also how often they escalate to high-casualty or mass-casualty levels, and how their distributions deviate from standard assumptions.

6.3 Graphical Representation





7 Milestone 7 – Simple Linear Regression and Correlation

7.1 Justification for Pairwise Iteration in Simple Linear Regression

In this milestone, I examined how individual attack characteristics relate linearly to the number of fatalities `nkill`. `nkill` was used as the single dependent variable, representing the outcome to be explained. The following variables were treated as independent predictors: `nwound`, `nperps`, `success`, `suicide`, `extended`, and `imonth`.

A series of simple linear regressions was run, each using one independent variable at a time against `nkill`. This produced pairs such as (`nwound`, `nkill`) and (`nperps`, `nkill`). This pairwise approach isolates the linear effect of each single characteristic on fatalities, without confounding from other predictors, and provides a clear view of which variables show meaningful linear association with `nkill`.

7.2 Analysis and Reflection

This analysis fit one simple linear regression model per independent variable to quantify correlation strength, R^2, and the regression line between that variable and `nkill`.

7.2.1 `nkill` vs `nwound` (Number Wounded)

- Pearson correlation $r = 0.5344$.
- $R^2 = 0.2856$.
- Regression equation: $\hat{Y} = 1.6705 + 0.1456X$.

This is the strongest linear relationship among all tested pairs: about 28.56% of the variance in `nkill` is explained by `nwound`. As the number wounded increases, the number killed also tends to increase, which matches the intuition that more severe attacks produce both more injuries and more deaths.

7.2.2 `nkill` vs `nperps` (Number of Perpetrators)

- Pearson correlation $r = 0.0250$.
- $R^2 = 0.0006$.
- Regression equation: $\hat{Y} = 2.5545 + 0.0015X$.

The relationship is extremely weak: less than 0.1% of the variance in `nkill` is explained by the number of perpetrators. Simply knowing how many perpetrators were involved has almost no linear predictive value for fatalities.

7.2.3 `nkill` vs `success` (Attack Success)

- Pearson correlation $r = 0.0531$.
- $R^2 = 0.0028$.
- Regression equation: $\hat{Y} = 0.6995 + 1.9245X$.

This model shows a very weak positive correlation: only 0.28% of the variance in `nkill` is explained by the success indicator. Although successful attacks may be more likely to have casualties, treating `success` (0/1) as a single linear predictor does not capture much of the variation in fatalities.

7.2.4 `nkill` vs `suicide` (Suicide Attack)

- Pearson correlation $r = 0.1361$.
- $R^2 = 0.0185$.
- Regression equation: $\hat{Y} = 2.0907 + 8.2102X$.

There is a weak positive correlation: about 1.85% of the variance in `nkill` is explained by whether an attack is a suicide attack. While suicide attacks are often associated with higher casualties, the linear model with this single binary predictor still explains only a small portion of the variability in deaths.

7.2.5 `nkill` vs `extended` (Extended Duration)

- Pearson correlation $r = 0.0277$.
- $R^2 = 0.0008$.
- Regression equation: $\hat{Y} = 2.3417 + 1.7269X$.

This pair also shows a very weak positive correlation, explaining less than 0.1% of the variance in `nkill`. Whether an attack is extended or not has negligible linear predictive power for fatalities when considered alone.

7.2.6 `nkill` vs `imonth` (Month of Attack)

- Pearson correlation $r = 0.0035$.
- $R^2 = 0.0000$.
- Regression equation: $\hat{Y} = 2.3268 + 0.0118X$.

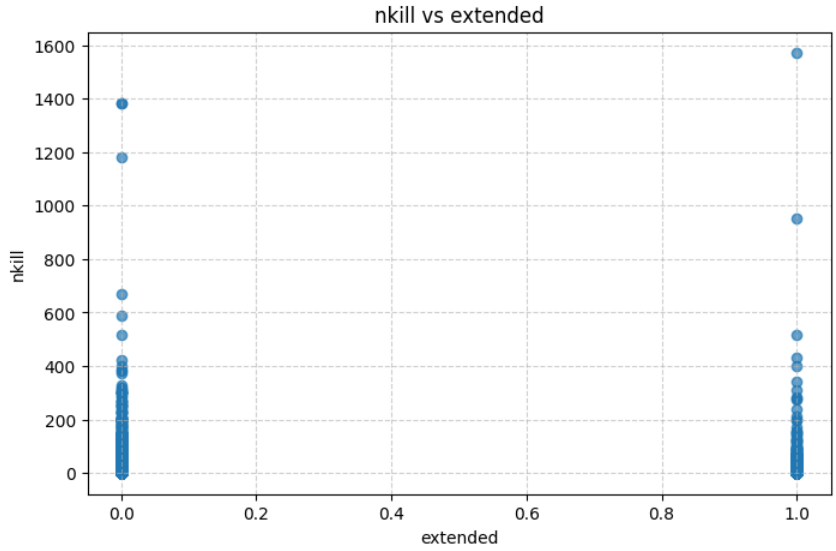
This relationship is essentially non-existent: the month in which an attack occurs explains almost none of the variance in the number of deaths, confirming the absence of a meaningful linear seasonal effect on fatalities.

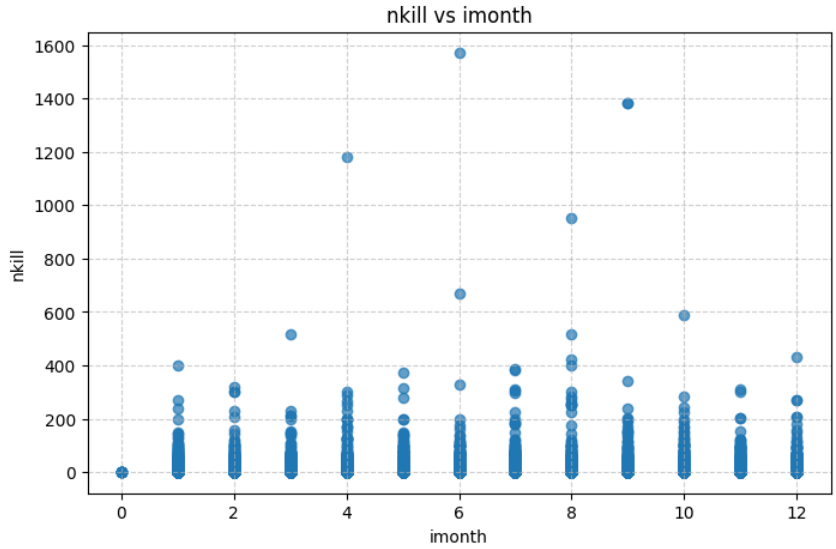
7.2.7 Comparison and Overall Reflection

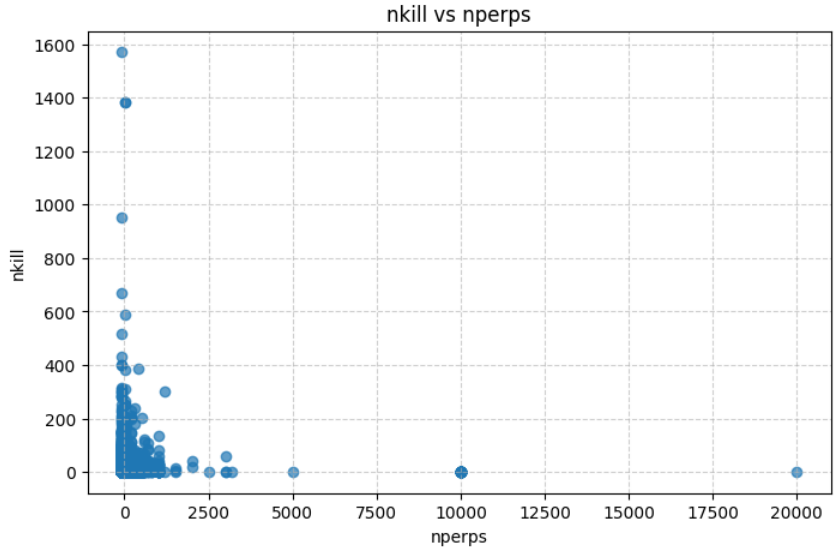
Comparing the R^2 values, `nwound` clearly stands out as the best single linear predictor of `nkill`, with a moderate relationship. All other variables—`nperps`, `success`, `suicide`, `extended`, and `imonth`—show very weak to negligible linear relationships, each explaining less than about 2% of the variance in fatalities when used alone.

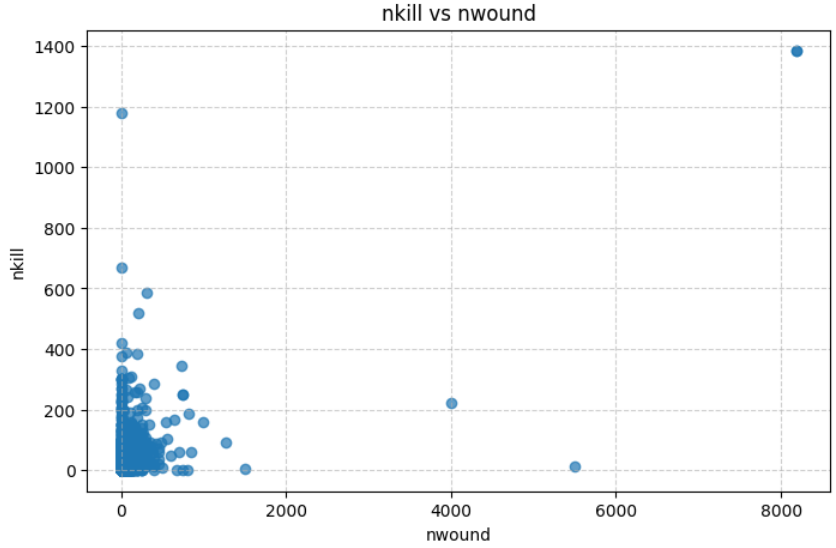
These results suggest that terrorism fatalities are driven by a complex combination of factors, and simple one-variable linear models are not sufficient to capture this complexity. `nwound` provides some insight into severity, but most variability in `nkill` remains unexplained at the single-predictor level. Future work would benefit from multiple regression, non-linear models, or other multivariate approaches that can incorporate target type, weapon type, context, and additional variables simultaneously. This milestone therefore serves as a baseline: it identifies `nwound` as a key correlate, but also emphasizes the inherently multivariate nature of fatality outcomes.

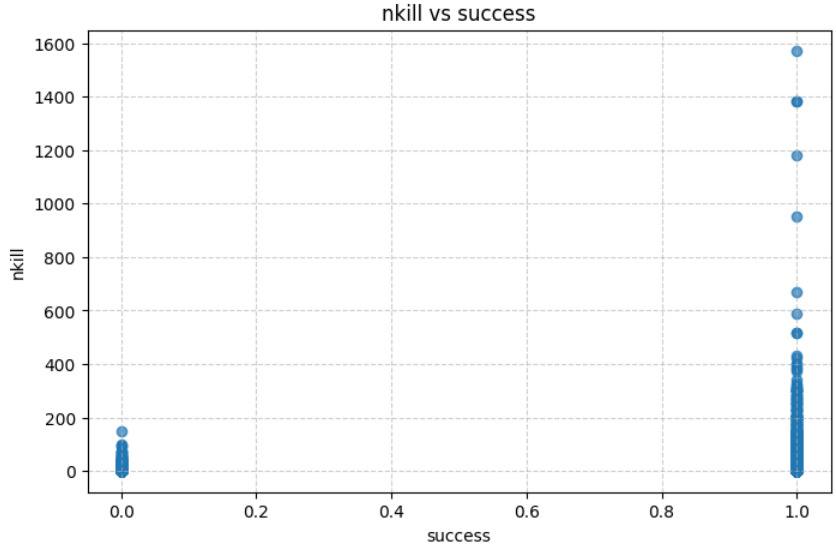
7.3 Graphical Representation (Partial)

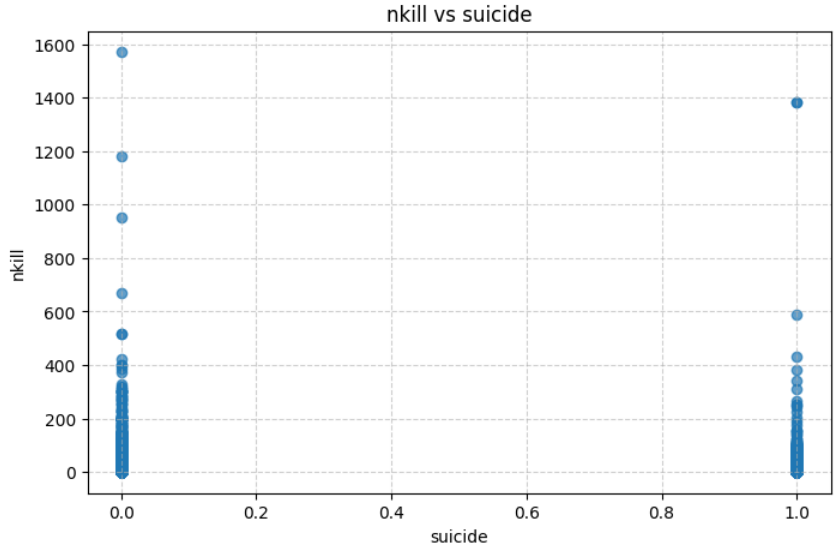


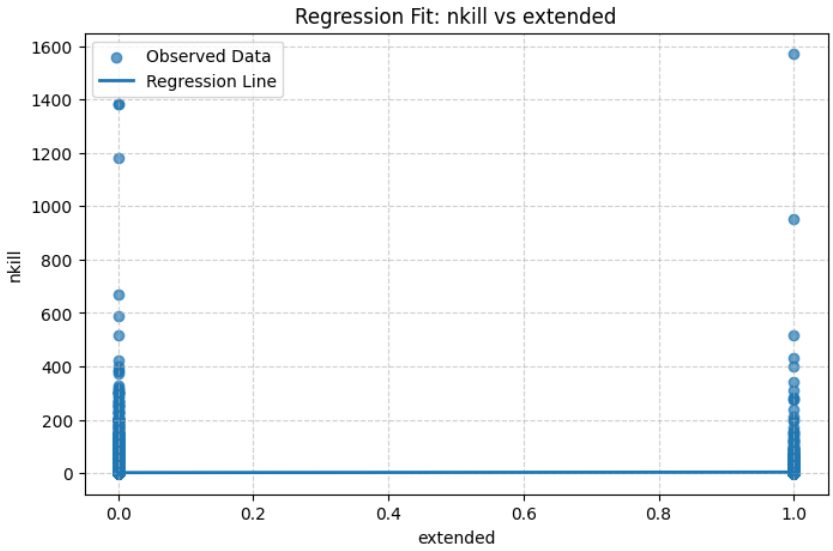


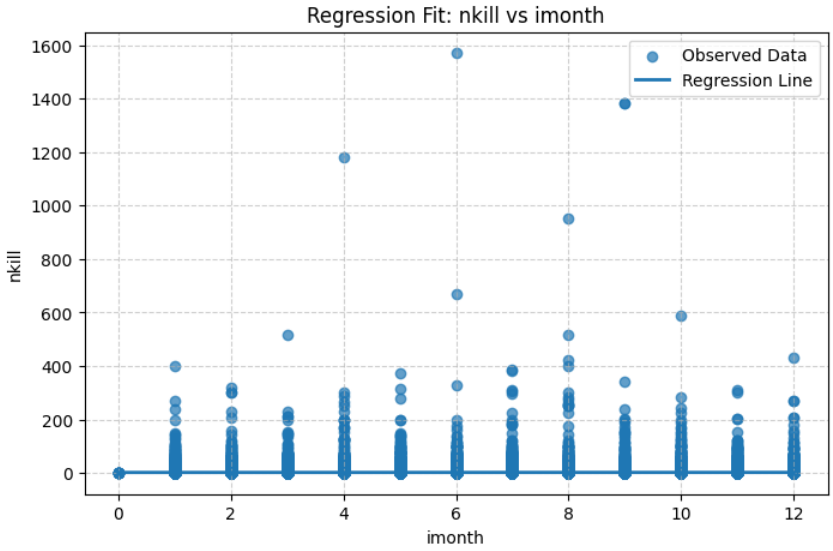


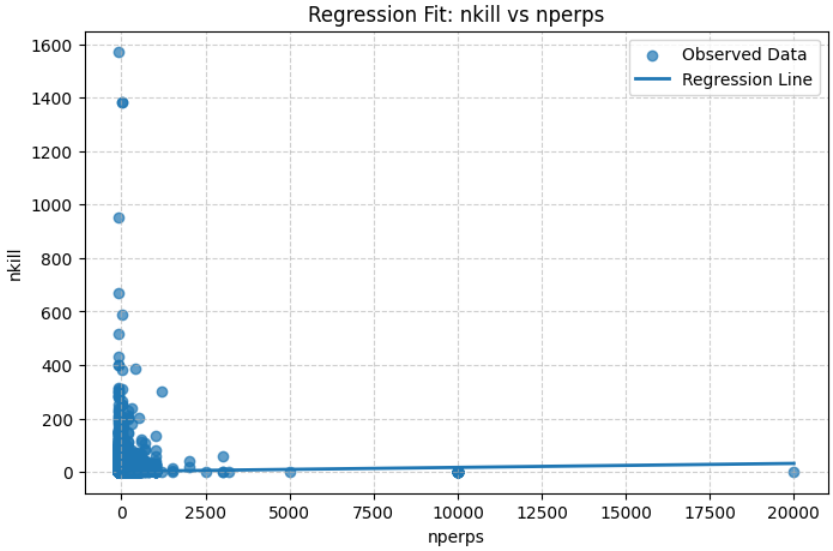

nkill vs nwound
nkill
nwound
0
200
400
600
800
1000
1200
1400
0
2000
4000
6000
8000


nkill vs success
nkill
success
0
200
400
600
800
1000
1200
1400
1600
0.0
0.2
0.4
0.6
0.8
1.0

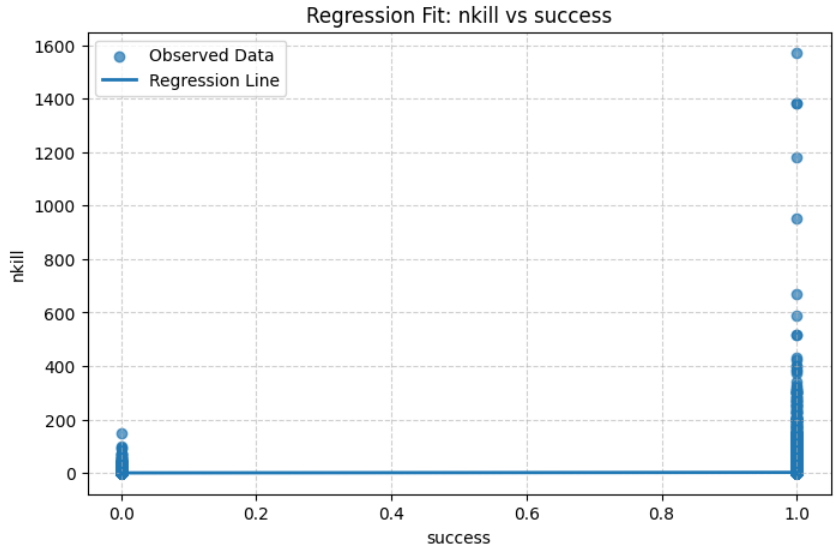

nkill vs suicide
nkill
suicide
0
200
400
600
800
1000
1200
1400
1600
0.0
0.2
0.4
0.6
0.8
1.0

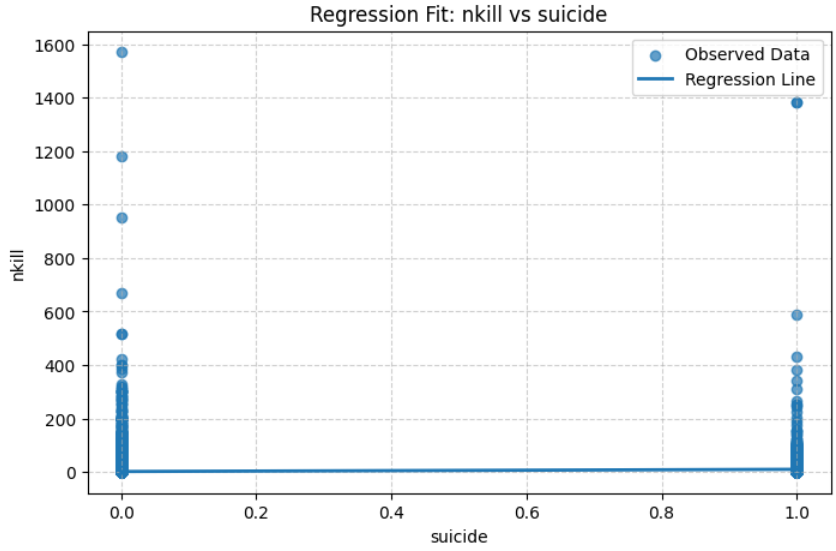

Regression Fit: nkill vs extended
Observed Data
Regression Line
nkill
extended
0
200
400
600
800
1000
1200
1400
1600
0.0
0.2
0.4
0.6
0.8
1.0


Regression Fit: nkill vs imonth
Observed Data
Regression Line
nkill
imonth
0
200
400
600
800
1000
1200
1400
1600
0
2
4
6
8
10
12


Regression Fit: nkill vs nperps
Observed Data
Regression Line
nkill
nperps
0
200
400
600
800
1000
1200
1400
1600
0
2500
5000
7500
10000
12500
15000
17500
20000


Regression Fit: nkill vs nwound
Observed Data
Regression Line
nkill
nwound
0
200
400
600
800
1000
1200
1400
0
2000
4000
6000
8000


Regression Fit: nkill vs success
Observed Data
Regression Line
nkill
success
0
200
400
600
800
1000
1200
1400
1600
0.0
0.2
0.4
0.6
0.8
1.0


Regression Fit: nkill vs suicide
Observed Data
Regression Line
nkill
suicide
0
200
400
600
800
1000
1200
1400
1600
0.0
0.2
0.4
0.6
0.8
1.0

Final Conclusion

This study shows that terrorist attacks are frequent, unevenly distributed over regions and time, and highly varied in their impact. Sampling and frequency analyses reveal a rising trend in incidents and a strongly skewed casualty profile, where many attacks are low-impact but a few are extremely deadly. Descriptive statistics and probability-based events highlight that successful attacks are more likely to be fatal, and that mass-casualty incidents, while rare, account for a large share of deaths. At the same time, some operations appear to target property or duration rather than maximizing fatalities. Simple linear regressions confirm that wounded counts offer the only moderate linear signal for predicting fatalities, and that most other variables have weak individual effects. Together, these results argue for careful data handling and richer, multivariate models when using this kind of data to inform policy or risk assessment.