Brittney Gross

6.1 Sourcing Open Data

10/5/2023

**Global Terrorism Database  
Data Source**

**Summary and Source:** I will be working with the Global Terrorism Database data set on [Kaggle](https://www.kaggle.com/datasets/START-UMD/gtd). Terrorism is defined as “the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation.” This data set is put together by looking at unclassified media articles. This data set is maintained by researchers at the National Consortium for the Study of Terrorism and Response to Terrorism (START), which is headquartered at the University of Maryland. I have sent an email and am awaiting a reply to receive written permission to post my results per their requirements.

**Data Collection:** The Global Terrorism Database is open data. However, the University of Maryland is using external data (unclassified media articles) to form this dataset. Starting in 2012 they started combining automated and manual data collection strategies. They apply a customized filter to media articles which include English articles and translations into English. Then the articles are put through natural language processing and machine learning to further refine the results. Finally, the University then manually reviews the articles to see which meet the inclusion criteria.

In a way, this data is administrative. Administrative data is data about administrative operations. While this data does not describe the operations of the University, they are using this data for their operation which is START. This is not usage data because they are not looking at who is using this data or for what. The data is also not interviews or surveys because the data was taken from media articles.

I chose this data set because I am interested in criminal justice and sociology. Obviously, this data set deals with crime/terrorism. If there was more time and datasets available, it would be interesting to bring in and analyze aspects of sociology.

**Limitations and Ethics:** There are several limitations to this data set. First, data from 1970-1997 were handwritten. This information had to be digitized. This could potentially lead to loss of data or errors. An example of this is that all the data from 1993 was lost and is not included in this data set. Second, starting in 2012 the University started combining automated and manual data collection strategies. This led to an influx of total number of worldwide terrorist attacks over 2011. This may lead to biases when trying to compare attacks overtime. This occurred due to an improved efficiency of the data collection process. Third, to avoid erroneous reporting the data is collected on a lag. This may provide more accurate data, but it may not be up to date. Fourth, the data comes from media articles. While The University includes media reports in that are in at least one high-quality source, not all countries have high-quality sources. In the same way, some countries may over or under report terrorism. This may lead to bias results.

In terms of ethics, there could be some bias in the data. For instance, people may have an implicit bias towards certain countries/groups of people/etc and consider one thing a terrorist attack in one country, but not the other. There could also be collection bias if some countries do not report terrorism. Thus, making it look as if certain countries do not have many terrorist attacks, but in reality, they are just not reported.

**Questions to Answer**:

1. What is the trend for terrorist attack over years for the entire world?
2. What is the rate of successful terrorist attacks over years for the entire world?
3. What is the terrorist attack trend over years per continent?
4. What is the success rate trend of terrorist attacks over years per continent?
5. What is the terrorist attack trend over years per country for the top 10 countries?
6. What is the successful terrorist attack trend over years per country for the top 10 countries?
7. What months have the most terrorist attacks?
8. What is the trend of successful terrorist attack over months?
9. What is the most common type of terrorist attack?
10. Do the terrorist attacks differ per to 10 countries?
11. Countries with percentage of terrorist attack per crit1, attacktype1\_txt, targtype1\_txt, weaptype1\_txt

**Global Terrorism Database  
Data Profile**

**Data Profile:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Description | Time Variant/Invariant | Structured/Unstructed | Qualitative/Quantitative | Nominal/Ordinal OR Discrete/Continuous |
| Iyear | Year of incident | Invariant | Structed | Quantitative | Continuous |
| imonth | Month of incident | Invariant | Structed | Quantitative | Continuous |
| Iday | Day of incident | Invariant | Structed | Quantitative | Continuous |
| extended | Duration of incident extended more/less than 24 hours | Invariant | Structed | Qualitative | Nominal |
| country\_txt | Country where incident occurred | Invariant | Structed | Qualitative | Nominal |
| region\_txt | Region where incident occurred | Invariant | Structed | Qualitative | Nominal |
| latitude | Latitude where incident occurred | Invariant | Structed | Quantitative | Continuous |
| longitude | Longitude where incident occurred | Invariant | Structed | Quantitative | Continuous |
| crit1 | The act was aimed at a political, economic, religious, or social goal | Invariant | Structed | Qualitative | Nominal |
| crit2 | The act was intended to coerce, intimidate, or covey another message to a larger audience | Invariant | Structed | Qualitative | Nominal |
| crit3 | The act is outside the context of legitimate warfare activities | Invariant | Structed | Qualitative | Nominal |
| doubtterr | Is there doubt whether an incident meets criteria for inclusion. | Invariant | Structed | Qualitative | Nominal |
| Multiple | Where several attacks are connected | Invariant | Structed | Qualitative | Nominal |
| Success | Was the attack successful | Invariant | Structed | Qualitative | Nominal |
| Suicide | Was the attack a suicide attack | Invariant | Structed | Qualitative | Nominal |
| attacktype1\_txt | General mode of attack text | Invariant | Structed | Qualitative | Nominal |
| targtype1\_txt | General type of target/victim text | Invariant | Structed | Qualitative | Nominal |
| targsubtype1 | Specific target category | Invariant | Structed | Qualitative | Nominal |
| targsubtype1\_txt | Specific target category text | Invariant | Structed | Qualitative | Nominal |
| natlty1\_txt | Nationality of the target that was attacked text | Invariant | Structed | Qualitative | Nominal |
| Gname | Name of group that carried out attack | Invariant | Unstructed | Qualitative | Nominal |
| Motive | Motive for attack | Invariant | Unstructed | Qualitative | Nominal |
| Gunceratin1 | Is the information about the name based on speculation | Invariant | Structed | Qualitative | Nominal |
| individual | Was the individual known to be affiliated with a group | Invariant | Structed | Qualitative | Nominal |
| nperps | Total number of terrorists participating in incident | Invariant | Structed | Quantitative | Discrete |
| nperpcap | Number of perpetrators taken into custody | Invariant | Structed | Quantitative | Discrete |
| claimed | Did the group claim responsibility for the attack | Invariant | Structed | Qualitative | Nominal |
| weaptype1\_txt | Main types of weapons text | Invariant | Structed | Qualitative | Nominal |
| weapsubtype1\_txt | Specific value for weapon text | Invariant | Structed | Qualitative | Nominal |
| nkill | Total number of fatalities | Invariant | Structed | Quantitative | Discrete |
| nkillus | Number of US fatalities | Invariant | Structed | Quantitative | Discrete |
| nkillter | Number of perpetrator fatalities | Invariant | Structed | Quantitative | Discrete |
| nwound | Total number of injured | Invariant | Structed | Quantitative | Discrete |
| nwoundus | Number of US injured | Invariant | Structed | Quantitative | Discrete |
| nwoundte | Number of perpetrators injured | Invariant | Structed | Quantitative | Discrete |
| property | Was there property damage | Invariant | Structed | Qualitative | Nominal |
| propextent\_txt | Extent of property damage text | Invariant | Structed | Qualitative | Nominal |
| ishostkid | Were there hostage or kidnapping victims | Invariant | Structed | Qualitative | Nominal |
| nhostkid | Total number of hostages/kidnapping victims | Invariant | Structed | Quantitative | Discrete |
| nhostkidus | Number of US hostages/kidnapping victims | Invariant | Structed | Quantitative | Discrete |
| hostkidoutcome\_txt | What was the outcome of the hostage/kidnapping text | Invariant | Structed | Qualitative | Nominal |
| nreleased | Number released/escaped/rescued | Invariant | Structed | Quantitative | Discrete |
| INT\_LOG | Location of attack vs nationality of the perpetrator group | Invariant | Structed | Qualitative | Nominal |
| INT\_IDEO | Nationality of perpetrator group and nationality of the target/victim(s) | Invariant | Structed | Qualitative | Nominal |
| INT\_MISC | Location of the attacks and the nationality of the target/victim | Invariant | Structed | Qualitative | Nominal |
| INT\_ANY | Location of attack: international vs domestic | Invariant | Structed | Qualitative | Nominal |

**Data Wrangling:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Columns dropped | Rows Dropped | Columns Renamed | Column's type changed | Comment/Reason |
|  | Deleted data prior to 2000 from iyear |  |  | too old of data |
| approxdate |  |  |  | too many missing variables |
| location |  |  |  | Redundant from lat and long |
| attacktype2 |  |  |  | too many missing variables |
| attacktype2\_txt |  |  |  | too many missing variables |
| attacktype3 |  |  |  | too many missing variables |
| attacktype3\_txt |  |  |  | too many missing variables |
| corp1 |  |  |  | unstructured data; more variability |
| target1 |  |  |  | unstructured data; more variability |
| targtype2 |  |  |  | too many missing variables |
| targtype2\_txt |  |  |  | too many missing variables |
| targsubtype2 |  |  |  | too many missing variables |
| targsubtype2\_txt |  |  |  | too many missing variables |
| corp2 |  |  |  | too many missing variables |
| target2 |  |  |  | too many missing variables |
| natlty2 |  |  |  | too many missing variables |
| natlty2\_txt |  |  |  | too many missing variables |
| targtype3 |  |  |  | too many missing variables |
| targtype3\_txt |  |  |  | too many missing variables |
| targsubtype3 |  |  |  | too many missing variables |
| targsubtype3\_txt |  |  |  | too many missing variables |
| corp3 |  |  |  | too many missing variables |
| target3 |  |  |  | too many missing variables |
| natlty3 |  |  |  | too many missing variables |
| natlty3\_txt |  |  |  | too many missing variables |
| gsubname |  |  |  | too many missing variables |
| gname2 |  |  |  | too many missing variables |
| gsubname2 |  |  |  | too many missing variables |
| gname3 |  |  |  | too many missing variables |
| gsubname3 |  |  |  | too many missing variables |
| guncertain2 |  |  |  | too many missing variables |
| guncertain3 |  |  |  | too many missing variables |
| claimmode |  |  |  | too many missing variables |
| claimmode\_txt |  |  |  | too many missing variables |
| claim2 |  |  |  | too many missing variables |
| claimmode2 |  |  |  | too many missing variables |
| claimmode2\_txt |  |  |  | too many missing variables |
| claim3 |  |  |  | too many missing variables |
| claimmode3 |  |  |  | too many missing variables |
| claimmode3\_txt |  |  |  | too many missing variables |
| compclaim |  |  |  | too many missing variables |
| weaptype2 |  |  |  | too many missing variables |
| weaptype2\_txt |  |  |  | too many missing variables |
| weapsubtype2 |  |  |  | too many missing variables |
| weapsubtype2\_txt |  |  |  | too many missing variables |
| weaptype3 |  |  |  | too many missing variables |
| weaptype3\_txt |  |  |  | too many missing variables |
| weapsubtype3 |  |  |  | too many missing variables |
| weapsubtype3\_txt |  |  |  | too many missing variables |
| weaptype4 |  |  |  | too many missing variables |
| weaptype4\_txt |  |  |  | too many missing variables |
| weapsubtype4 |  |  |  | too many missing variables |
| weapsubtype4\_txt |  |  |  | too many missing variables |
| weapdetail |  |  |  | unstructured data; more variability |
| propvalue |  |  |  | too many missing variables |
| propcomment |  |  |  | too many missing variables |
| nhours |  |  |  | too many missing variables |
| ndays |  |  |  | too many missing variables |
| divert |  |  |  | too many missing variables |
| kidhijcountry |  |  |  | too many missing variables |
| ransom |  |  |  | too many missing variables |
| ransomamt |  |  |  | too many missing variables |
| ransomamtus |  |  |  | too many missing variables |
| ransompaid |  |  |  | too many missing variables |
| ransompaidus |  |  |  | too many missing variables |
| ransomnote |  |  |  | too many missing variables |
| addnotes |  |  |  | unstructured data; more variability |
| scite1 |  |  |  | not relevant to analysis |
| scite2 |  |  |  | not relevant to analysis |
| scite3 |  |  |  | not relevant to analysis |
| dbsource |  |  |  | not relevant to analysis |
| related |  |  |  | too many missing variables |
| country |  |  |  | redundant from country\_txt |
| region |  |  |  | redundant from region\_txt |
| eventid |  |  |  | redundant from date; not unique identifier |
| resolution |  |  |  | too many missing variables |
| provstate |  |  |  | redundant from latitude and longitude |
| city |  |  |  | redundant from latitude and longitude |
| specificity |  |  |  | redundant from latitude and longitude |
| vicinity |  |  |  | redundant from latitude and longitude |
| summary |  |  |  | not relevant to analysis |
| alternative |  |  |  | not relevant to analysis |
| alternative\_txt |  |  |  | not relevant to analysis |
| attacktype1 |  |  |  | redundant from attacktype1\_txt |
| targtype1 |  |  |  | redundant from targtype1\_txt |
| targsubtype1 |  |  |  | redundant from targsubtype1\_txt |
| natlty1 |  |  |  | redundant from natlty1\_txt |
| weaptype1 |  |  |  | redundant from weaptype1\_txt |
| weapsubtype1 |  |  |  | redundant from weapsubtype1\_txt |
| propextent |  |  |  | redundant from propextent\_txt |
| hostkidoutcome |  |  |  | redundant from hostkidoutcome\_txt |
|  |  |  | Extended | Changed to Category because this number respresents text, not a number |
|  |  |  | Crit1 | Changed to Category because this number respresents text, not a number |
|  |  |  | Crit2 | Changed to Category because this number respresents text, not a number |
|  |  |  | crit3 | Changed to Category because this number respresents text, not a number |
|  |  |  | doubtterr | Changed to Category because this number respresents text, not a number |
|  |  |  | multiple | Changed to Category because this number respresents text, not a number |
|  |  |  | success | Changed to Category because this number respresents text, not a number |
|  |  |  | suicide | Changed to Category because this number respresents text, not a number |
|  |  |  | guncertain1 | Changed to Category because this number respresents text, not a number |
|  |  |  | individual | Changed to Category because this number respresents text, not a number |
|  |  |  | claimed | Changed to Category because this number respresents text, not a number |
|  |  |  | property | Changed to Category because this number represents text, not a number |
|  |  |  | ishotkid | Changed to Category because this number represents text, not a number |
|  |  |  | int\_log | Changed to Category because this number represents text, not a number |
|  |  |  | int\_ideo | Changed to Category because this number represents text, not a number |
|  |  |  | int\_misc | Changed to Category because this number represents text, not a number |
|  |  |  | int\_any | Changed to Category because this number represents text, not a number |

**Data Cleaning/Consistency Checks:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Missing Values** | **Missing Values Treatment** | **Duplicates** |
| **Terrorism Data Set** |  |  | **3818 duplicates that were removed** |
| **Terrorism Data Set** | **807 missing latitude values** | **Will use country and region instead of lat/long** |  |
| **Terrorism Data Set** | **807 longitude values** | **Will use country and region instead of lat/long** |  |
| **Terrorism Data Set** | **338 guncertain1 values** | **Nothing since <1% of data is missing** |  |
| **Terrorism Data Set** | **10986 nperps values** | **Will not use these values in analysis; If I were going to, I would make mention that these values were not listed but not import the mean as it may throw off the results** |  |
| **Terrorism Data Set** | **2747 nperpcap** | **Will not use these values in analysis; If I were going to, I would make mention that these values were not listed but not import the mean as it may throw off the results** |  |
| **Terrorism Data Set** | **3873 nkill values** | **Will not use these values in analysis; If I were going to, I would make mention that these values were not listed but not import the mean as it may throw off the results** |  |
| **Terrorism Data Set** | **340 nkillus values** | **Will not use these values in analysis; If I were going to, I would make mention that these values were not listed but not import the mean as it may throw off the results** |  |
| **Terrorism Data Set** | **2430 nkillter values** | **Will not use these values in analysis; If I were going to, I would make mention that these values were not listed but not import the mean as it may throw off the results** |  |
| **Terrorism Data Set** | **7422 nwound values** | **Will not use these values in analysis; If I were going to, I would make mention that these values were not listed but not import the mean as it may throw off the results** |  |
| **Terrorism Data Set** | **539 nwoundus values** | **Will not use these values in analysis; If I were going to, I would make mention that these values were not listed but not import the mean as it may throw off the results** |  |
| **Terrorism Data Set** | **4101 nwoundte values** | **Will not use these values in analysis; If I were going to, I would make mention that these values were not listed but not import the mean as it may throw off the results** |  |
| **Terrorism Data Set** | **97262 nhostkid values** | **These column represents the total number of hostages. When looking at number of hostage incidents and comparing these missing values, there are less than 1% missing. So I will do nothing.** |  |
| **Terrorism Data Set** | **97315 nhostkidus values** | **These column represents the total number of US hostages. When looking at number of hostage incidents and comparing these missing values, there are less than 1% missing. So I will do nothing.** |  |
| **Terrorism Data Set** | **97497 nreleased values** | **These column represents the total number of hostages released. When looking at number of hostage incidents and comparing these missing values, there are less than 1% missing. So I will do nothing.** |  |

**Basic Descriptive Statistics:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **iyear** | **imonth** | **iday** | **latitude** | **longitude** | **nperps** | **nperpcap** | **nkill** | **nkillus** | **nkillter** | **nwound** | **nwoundus** | **nwoundte** | **nhostkid** | **nhostkidus** | **nreleased** |
| **count** | **106158** | **106158** | **106158** | **105350** | **105351** | **19019** | **102573** | **102285** | **105818** | **103728** | **98736** | **105619** | **102057** | **8463** | **8793** | **5377** |
| mean | **2012.182** | **6.465033** | **15.58617** | **26.37308** | **51.78824** | **13.97192** | **0.120617** | **2.603177** | **0.038528** | **0.52254** | **3.861206** | **0.01638** | **0.113172** | **11.38485** | **0.050836** | **6.831505** |
| std | **4.163704** | **3.383985** | **8.819229** | **13.30245** | **36.19208** | **87.67779** | **2.123492** | **12.61109** | **5.874611** | **4.23566** | **39.85097** | **0.757734** | **1.549536** | **90.5015** | **1.336063** | **56.39372** |
| min | **2000** | **1** | **0** | **-42.884** | **-124.225** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** |
| 25% | **2010** | **4** | **8** | **15.47004** | **39.35555** | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **1** | **0** | **0** |
| 50% | **2013** | **6** | **15** | **32.37001** | **44.84359** | **2** | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **2** | **0** | **1** |
| 75% | **2015** | **9** | **23** | **34.20842** | **71.0025** | **5** | **0** | **2** | **0** | **0** | **3** | **0** | **0** | **5** | **0** | **2** |
| max | **2017** | **12** | **31** | **65.82512** | **179.3667** | **5000** | **406** | **1570** | **1360** | **500** | **8191** | **151** | **200** | **5350** | **86** | **2769** |