**Final_Report.md**

# An Analysis of the Global Terrorism Database

## Introduction

It’s often been said that one should never negotiate with terrorists. However, in this age of abundant data and documented events, analysts can leverage their expertise along with data to assess how to deal with terrorists and react to attacks. One of the most comprehensive open-source and unclassified datasets of terrorist attacks is the Global Terrorism Database (GTD). It is available online so that terrorist violence can be studied and conquered by anyone. It has been maintained and has steadily grown by a team of researchers as well as technical staff. As the name suggests, terrorist attacks from all over the world are documented in this database. It currently holds over 200,000 incidents from 1970 to 2019. Each entry is meticulously documented with numerical and categorical data, ready for analysis with minimal cleaning needed. This database defines a terrorist attack as the threat or actual use of illegal force and violence by a non-state actor to attain political, economic, religious, or social goal through fear, coercion, or intimidation.

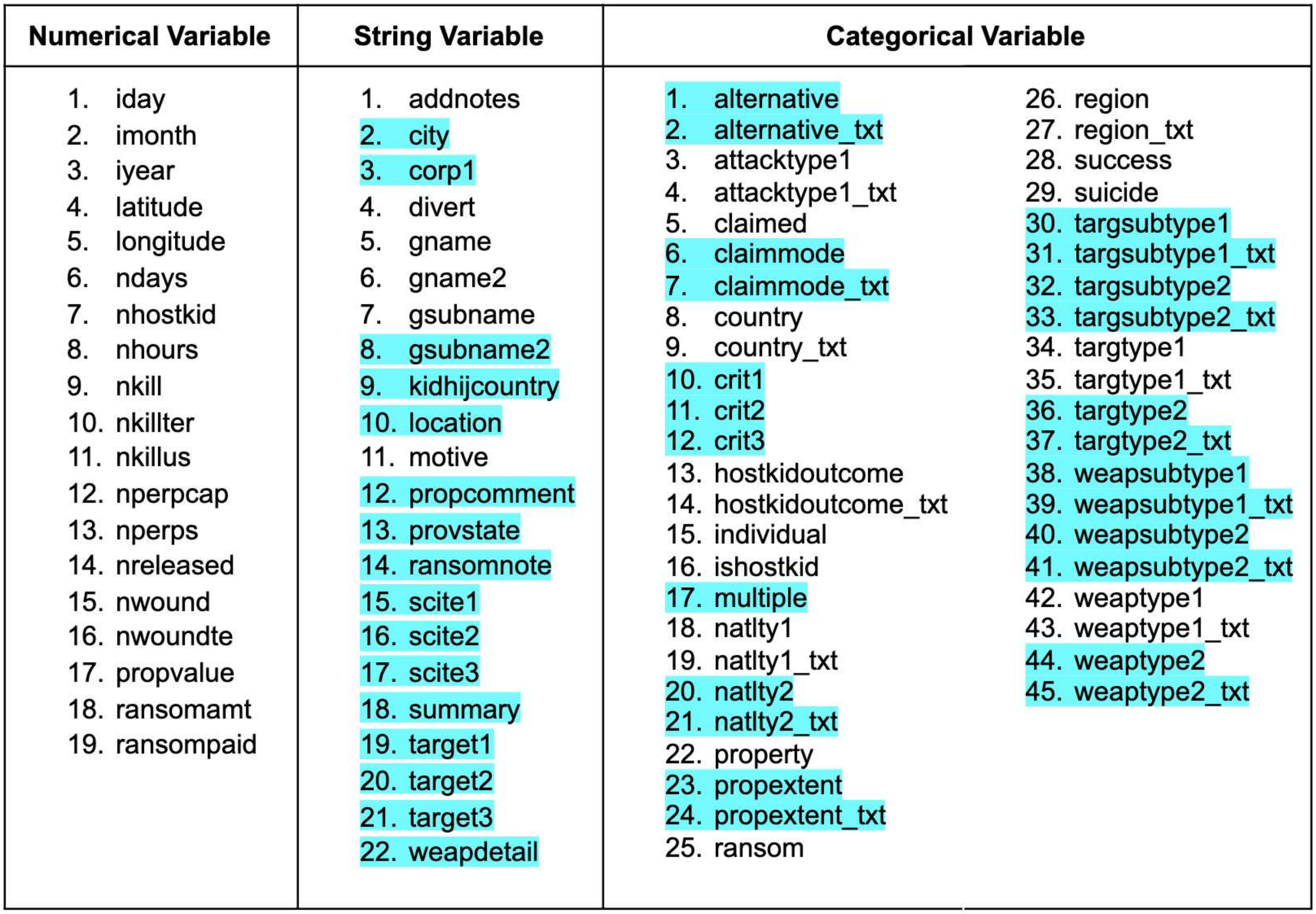
## Objective

The research team had one objective, which was to find scenarios that are more likely to have an event. The research completed can help drive future actions to mitigate, prepare for, or prevent terrorist like events from happening in the future. There were three areas we focused on in order to complete our objective. For purposes of this analysis, we limited our analysis to events that occurred from 2000 to 2019. The GTD researchers/data collectors noted that there were some data quality issues prior to 2000 as data collection methods have changed over the decades.

1. Time: The number of events have fluctuated over 50 years. We aggregated events over years all the way down to specific days of the week to discover patterns and trends.
2. Location: Certain areas of the world have experienced more events than others. We focus on event location to find hotspots in the world.
3. Motive: All terrorist-like events have a motive. We focused on common words found in the motive column for any patterns.
4. Attack Type and Impact: Every terrorist-like event has a means of getting it done. We focused on the most prevalent types of terrorist attack and the types that caused the most fatalities.

## The Dataset

Below is a table of the variables that were found in the Global Terrorism Database (GTD). The variables highlighted in blue were removed prior to analysis.



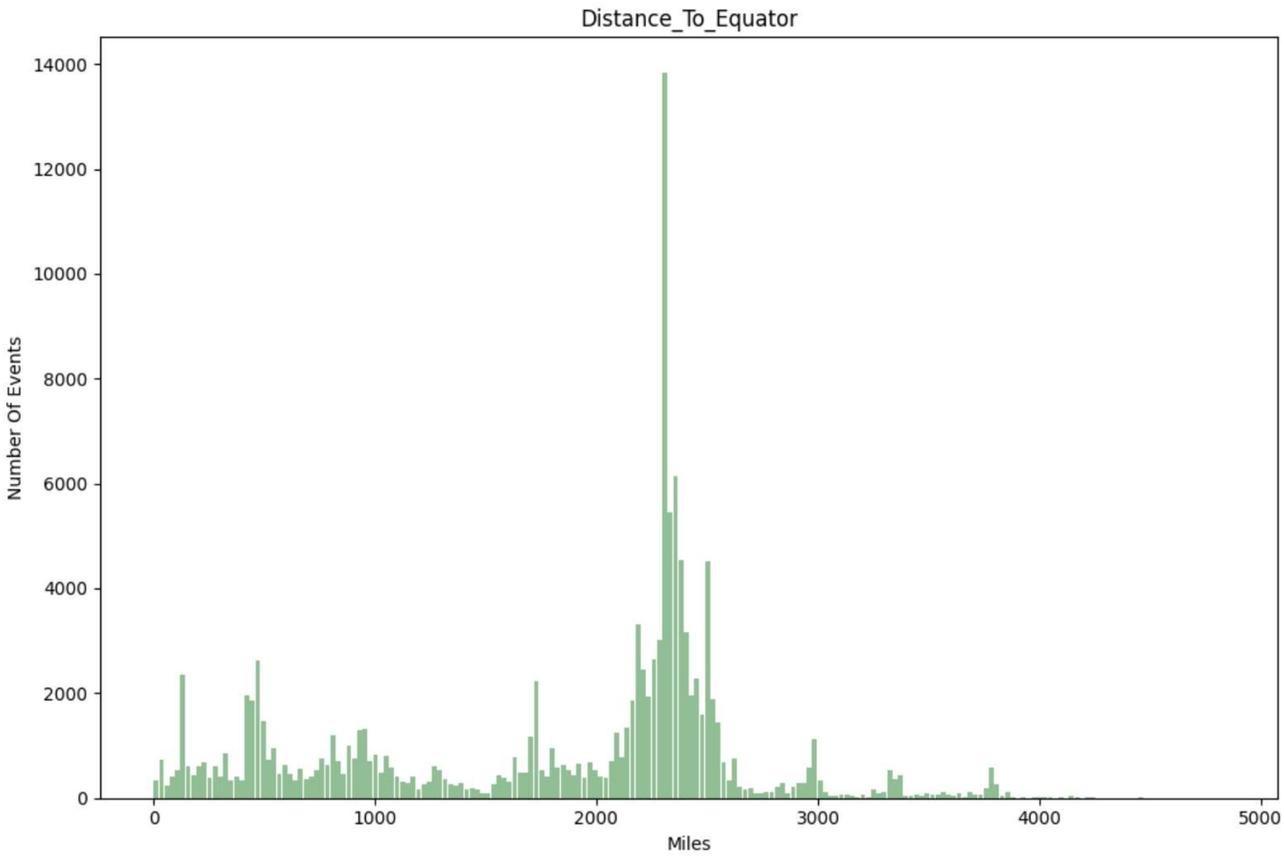
| Numerical Variable | String Variable | Categorical Variable |
|---|---|---|
| 1. iday<br>2. imonth<br>3. iyear<br>4. latitude<br>5. longitude<br>6. ndays<br>7. nhostkid<br>8. nhours<br>9. nkill<br>10. nkillter<br>11. nkillus<br>12. nperpcap<br>13. nperps<br>14. nreleased<br>15. nwound<br>16. nwoundte<br>17. propvalue<br>18. ransomamt<br>19. ransompaid | 1. addnotes<br>2. city<br>3. corp1<br>4. divert<br>5. gname<br>6. gname2<br>7. gsubname<br>8. gsubname2<br>9. kidhijcountry<br>10. location<br>11. motive<br>12. propcomment<br>13. provstate<br>14. ransomnote<br>15. scite1<br>16. scite2<br>17. scite3<br>18. summary<br>19. target1<br>20. target2<br>21. target3<br>22. weapdetail | 1. alternative<br>2. alternative_txt<br>3. attacktype1<br>4. attacktype1_txt<br>5. claimed<br>6. claimmode<br>7. claimmode_txt<br>8. country<br>9. country_txt<br>10. crit1<br>11. crit2<br>12. crit3<br>13. hostkidoutcome<br>14. hostkidoutcome_txt<br>15. individual<br>16. ishostkid<br>17. multiple<br>18. natlty1<br>19. natlty1_txt<br>20. natlty2<br>21. natlty2_txt<br>22. property<br>23. propextent<br>24. propextent_txt<br>25. ransom<br>26. region<br>27. region_txt<br>28. success<br>29. suicide<br>30. targsubtype1<br>31. targsubtype1_txt<br>32. targsubtype2<br>33. targsubtype2_txt<br>34. targtype1<br>35. targtype1_txt<br>36. targtype2<br>37. targtype2_txt<br>38. weapsubtype1<br>39. weapsubtype1_txt<br>40. weapsubtype2<br>41. weapsubtype2_txt<br>42. weaptype1<br>43. weaptype1_txt<br>44. weaptype2<br>45. weaptype2_txt |

## Where Did These Events Happen In The World?

The GTD contained the longitude and latitude over all events. The dataset also contained city name, but we felt it was more accurate to utilize the lat/long datapoints to gather more detailed information on the events, as well as combine exterior datasets to the GTD.

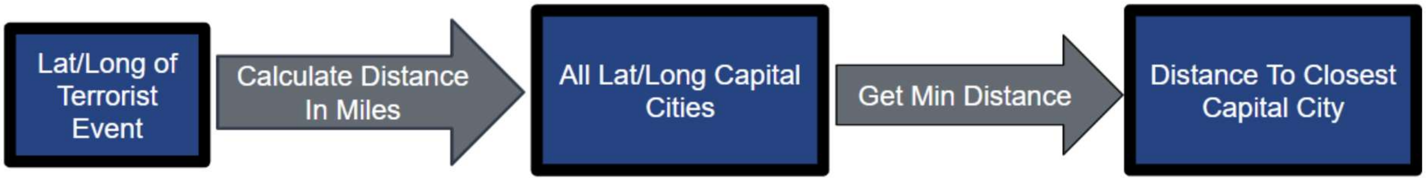
To calculate distances from events we used the Haversine formula, which calculates the distance between two points on a sphere. We added in the radius of the earth (in miles) to convert the function into the most realistic distance estimator for our research. This formula does not take into account elevation between two points on earth.

The first observation we analyzed was how close each event was to the equator. We noticed a left skewed distribution of events, with a large spike in the 2000 to 3000 mile range. This distance coincides with events that have occurred in the Middle East.



Our second observation using lat/long was calculating the distance between and event and the closest capital city. In order to do this we brought in detailed information of the worlds cities and focused solely on capital cities.

For each event we calculated the distance to all capital cities, and brought back the minimum distance.



Lat/Long of Terrorist Event

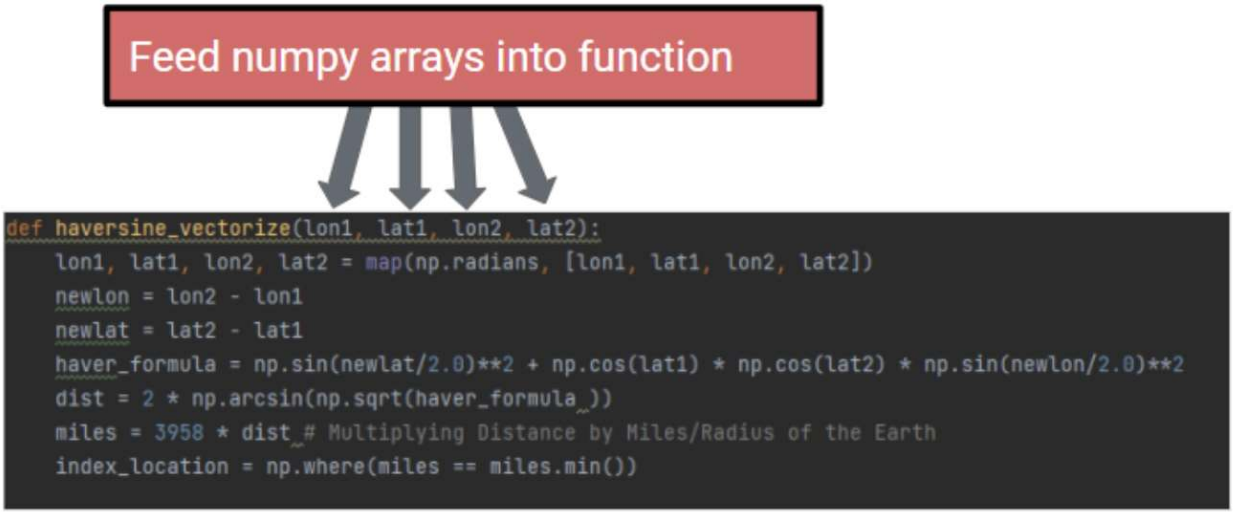
Calculate Distance In Miles

All Lat/Long Capital Cities

Get Min Distance

Distance To Closest Capital City

This allowed us to validate if an event occurred either inside or outside the capital city. We used the threshold of 20 miles from the center point of a city to determine if an event was within city bounds. In order to complete this computation intensive function across all events, we needed to utilize numpy arrays and vectorization in order to efficiently calculate all distances.



Feed numpy arrays into function

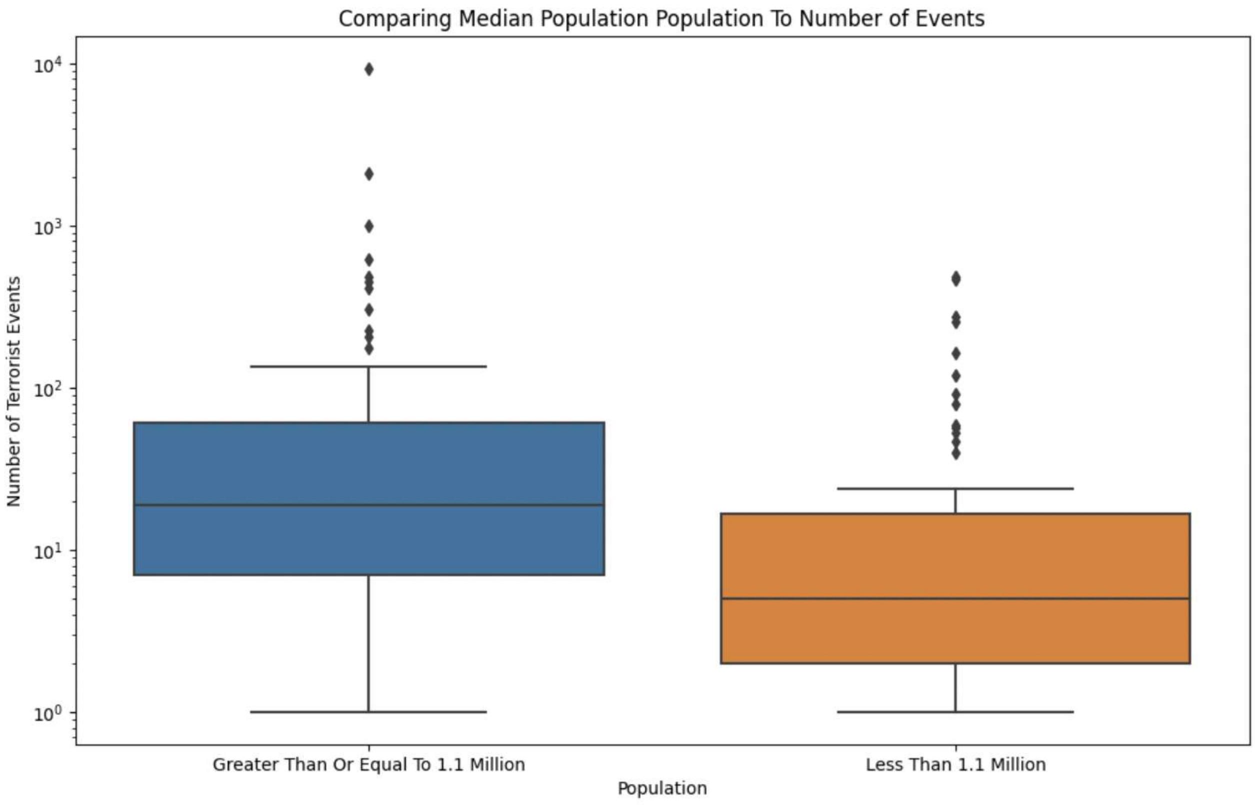
```
def haversine_vectorize(lon1, lat1, lon2, lat2):
    lon1, lat1, lon2, lat2 = map(np.radians, [lon1, lat1, lon2, lat2])
    newlon = lon2 - lon1
    newlat = lat2 - lat1
    haver_formula = np.sin(newlat/2.0)**2 + np.cos(lat1) * np.cos(lat2) * np.sin(newlon/2.0)**2
    dist = 2 * np.arcsin(np.sqrt(haver_formula ))
    miles = 3958 * dist # Multiplying Distance by Miles/Radius of the Earth
    index_location = np.where(miles == miles.min())
```

This allowed us to complete approximately 27 million calculations with a computing time of 10 minutes.

There are some fascinating results:

- ~15% of all terrorist events happen within a capital city
- The average distance of a capital city event is 5 miles from the center of the city

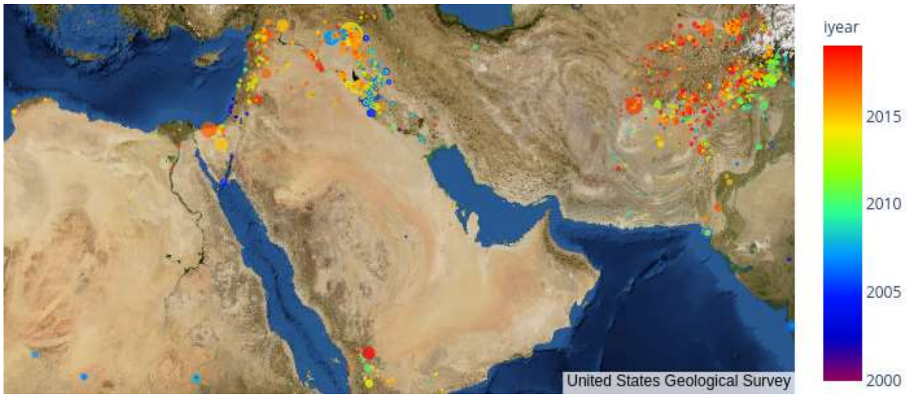
After identifying capital city events, we merged population data from the world cities dataset to find if there is a correlation between the population of a city, and the number of events. We split the dataset into two groups; cities that have a population greater than the median (1.1 million) and cities that have a population less than the median. You can see that on average, larger cities tend to have more events.



## Locations of events

When assessing where certain attacks happen and their impact in terms of the number of people killed, a choropleth map was the best visualization to use. To see how deaths occurred overtime, the primary data we decided to include was the latitudinal and longitudinal coordinates to plot on the map, and 5 additional informative variables which are in the hover box: country, year, terrorist group name, number of people wounded, and the primary target type. The visualization was created with plotly express which allows for many customizable options such as marker size and legend. For this visualization, we used the year for the color of each marker and the number of people killed in that attack for the marker size. This gave us the ability to quickly gloss over the map and see when different surges of attacks occurred across the globe, as well as how big their impact was in deaths.

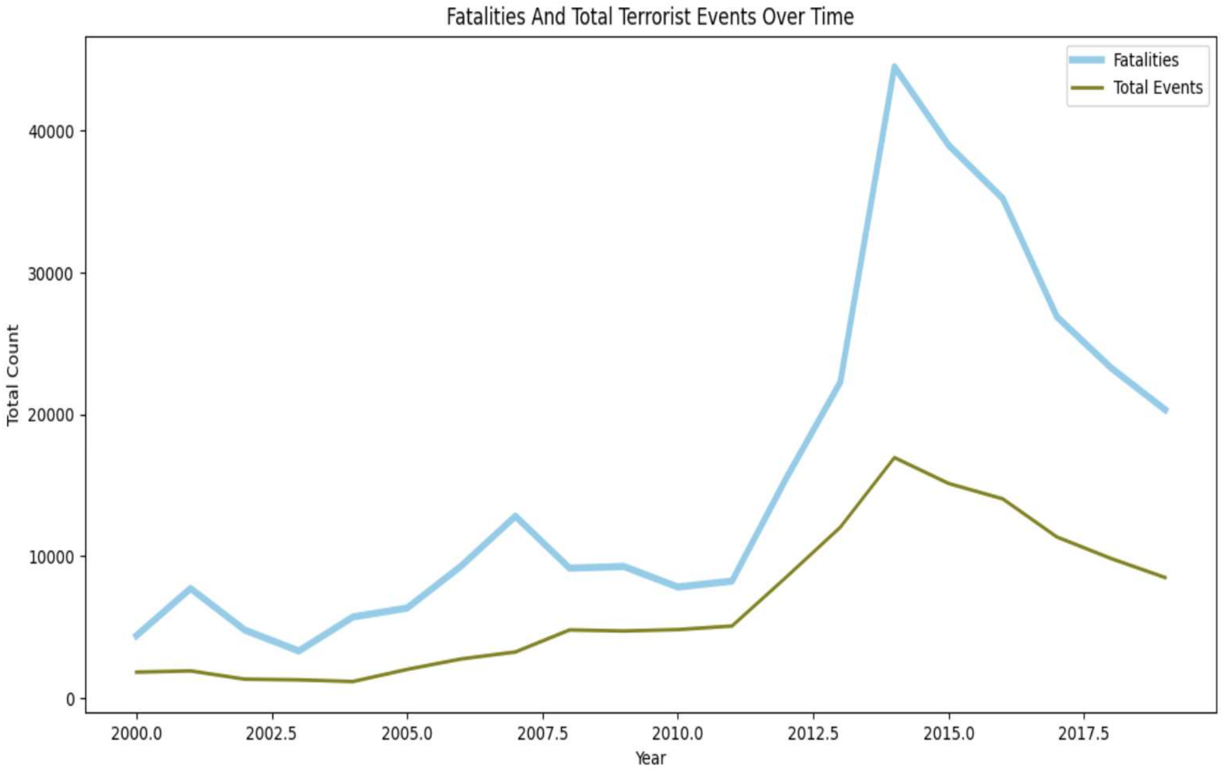
We did note that the attack on the World Trade Center on September 11, 2001 is an outlying marker worth focussing on. This is for 2 main reasons. Firstly, it simply stands out due to its sheer size relative to the other events. It is the largest attack by over 2 times the deaths of its runner up, an attack that occurred in Iraq by ISIL in 2014 where 670 were killed; on 9/11, 1,385 American lives were lost. Secondly, this marker stands out as it is one of the few large attacks in the West. It catches the eye with its purple color indicating it occurred in the early 2000s, while most other large and clustered markers are green, red, and yellow, indicating that they occurred in the last decade with many other attacks in the same year.



## Attacks Over Time

Looking at trends over time could help us discover more prominent time periods of events. Our strategy was to look at events from a high level point of view in regards to time, and step down to a more granular level, all the way to day of the week.

From a high level point of view we can see fluctuations of events over the past two decades.

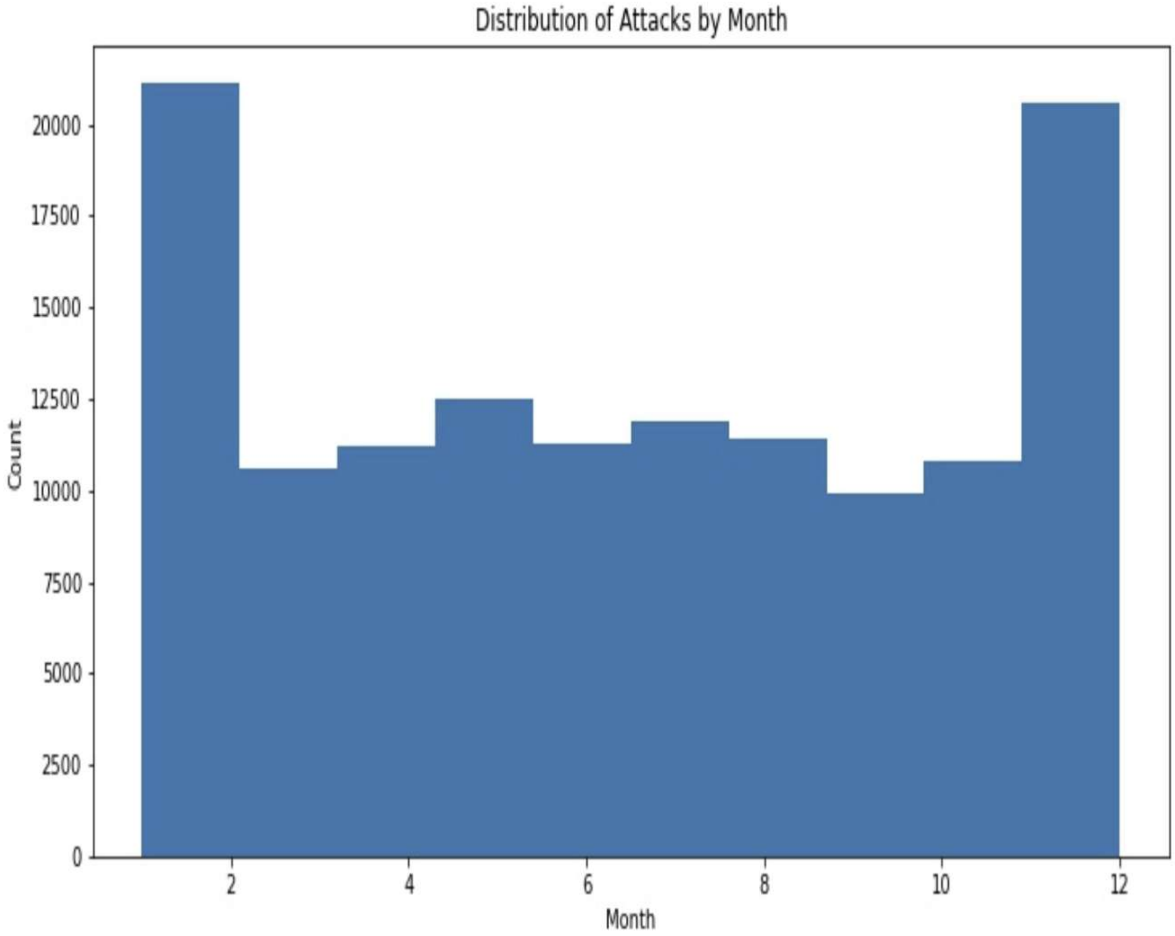


The number of fatalities coincides with the number of events per year, with 2014 being the peak number of events totally to around 40,000 fatalities. The number of events as well as fatalities has steadily decreased, but remains relatively high.

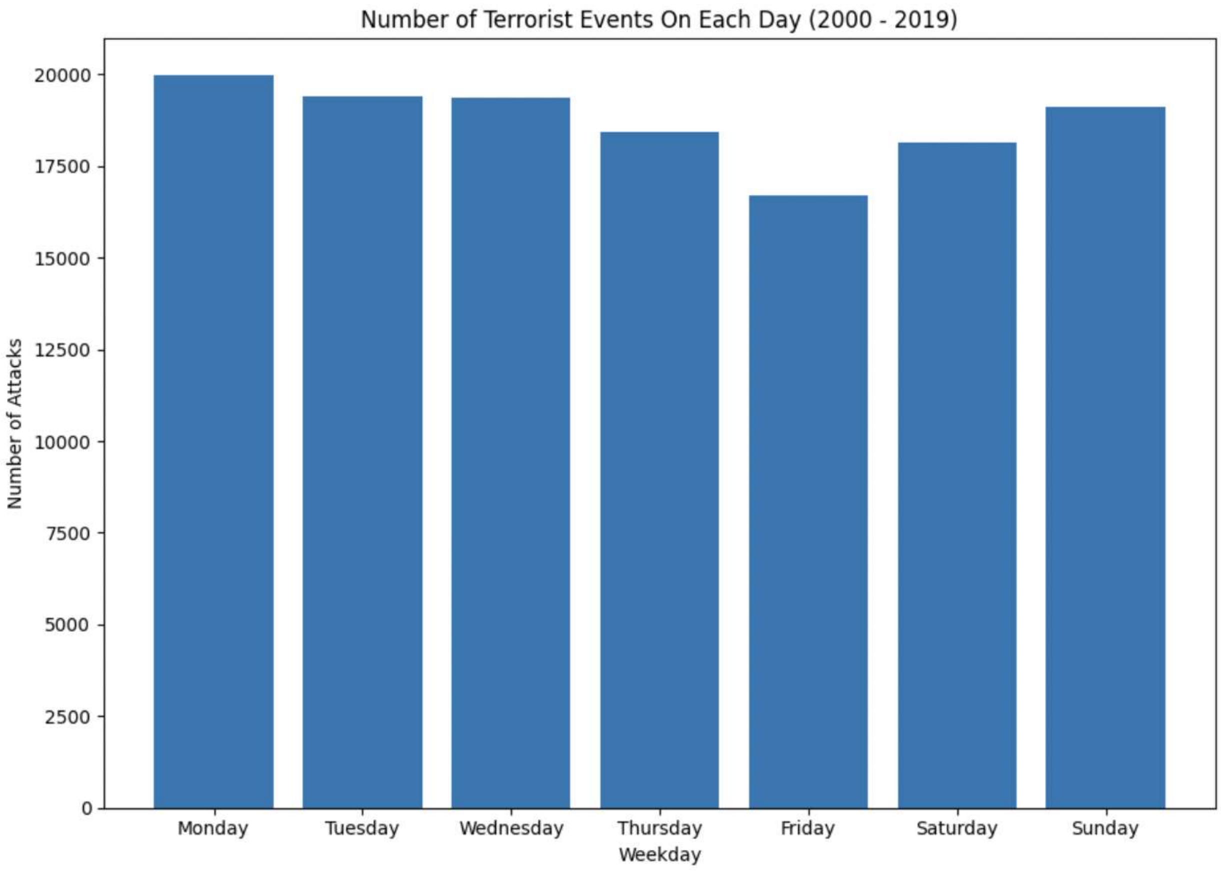
From a yearly point of view we can see the total number of events is substantially higher in January and December. As of right now it is unknown why these two months have substantially more, but a couple of plausible reasons could be made:

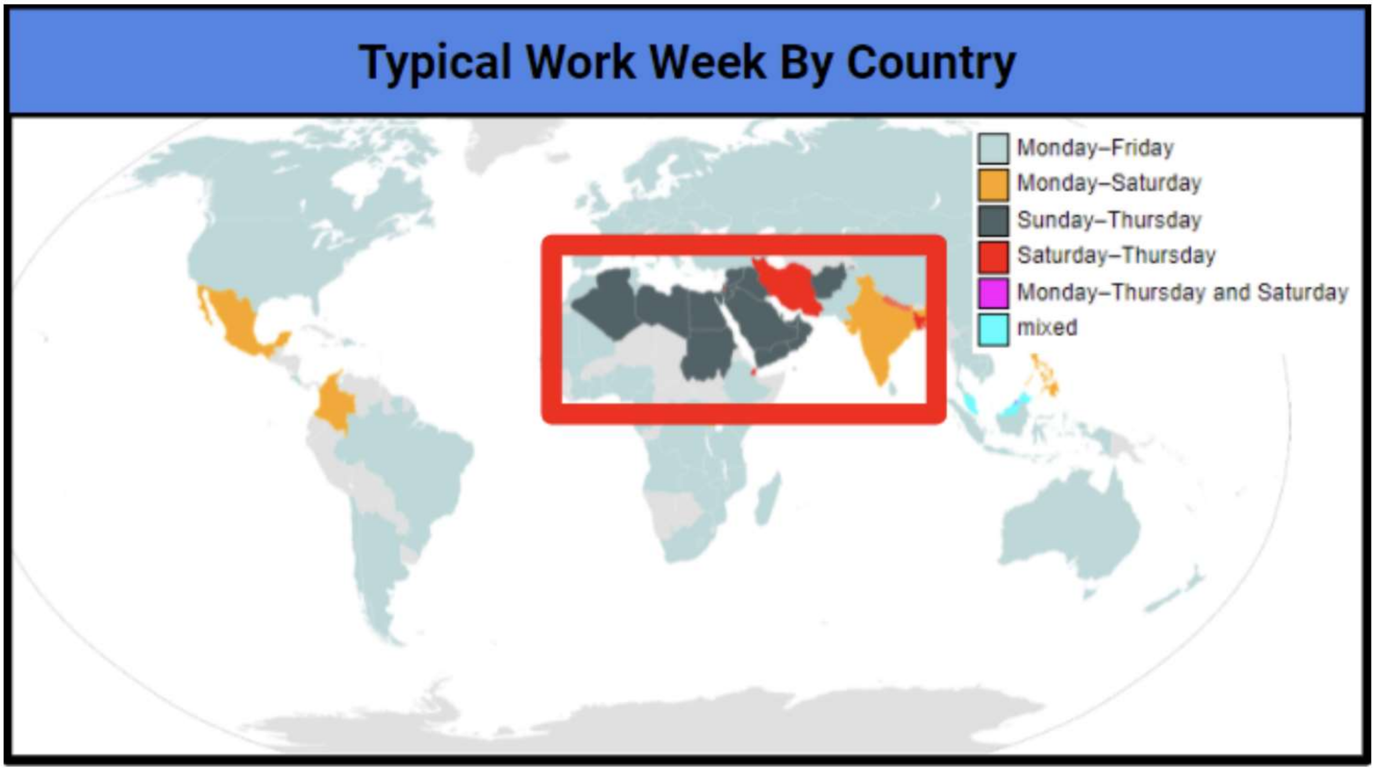
- Religious Holidays
- Extended Time off for citizens
- Election Cycle

Further research and additional datasets need to be analyzed to prove these assumptions.



Zooming in further, We see there are significant patterns when looking at the number of attacks on each day of the week. The number of attacks peaks on Monday and settles on Friday only to rise again. By both using the location of most events as well as the typical work week we can see a plausible reason why Fridays are the least affected by terrorist attacks.


Number of Terrorist Events On Each Day (2000 - 2019)
Number of Attacks
20000
17500
15000
12500
10000
7500
5000
2500
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday

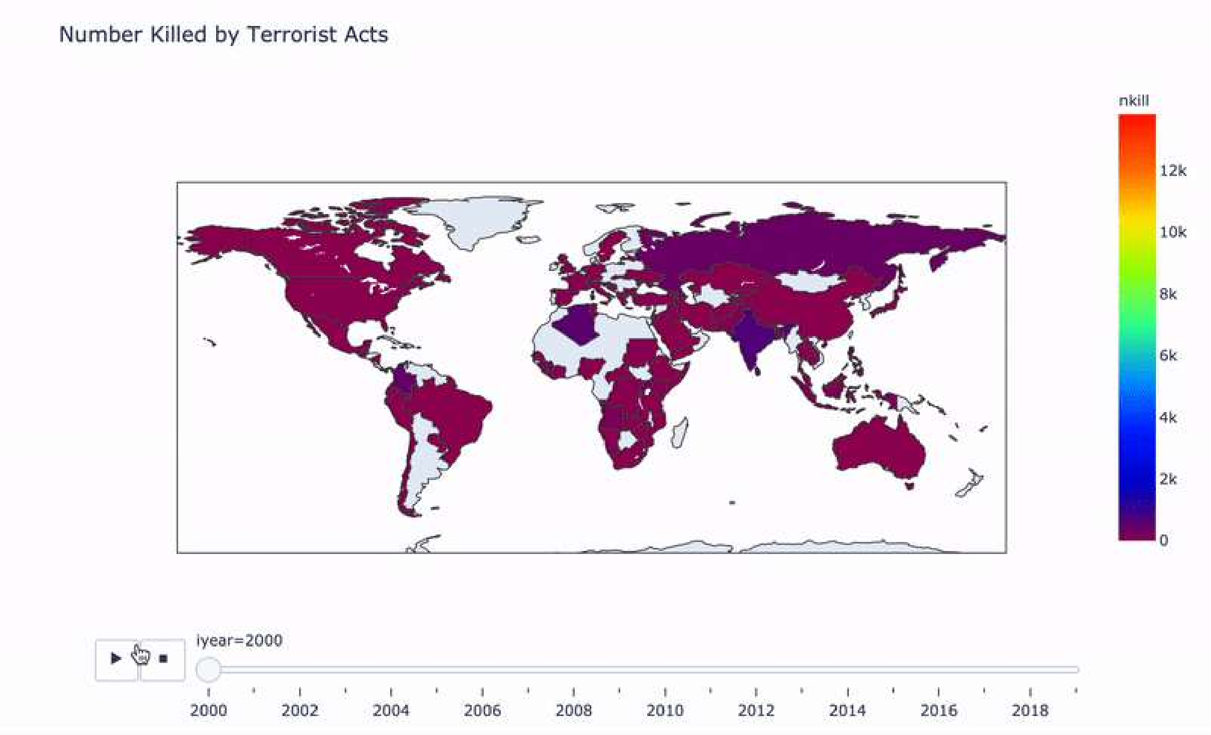

Typical Work Week By Country
Monday–Friday
Monday–Saturday
Sunday–Thursday
Saturday–Thursday
Monday–Thursday and Saturday
mixed

Most events have occured in the red box, typically in the Middle East, where a work week is usually Sunday through Thursday, or a Saturday through Thursday. This coincides with the fact that Fridays are the weekend, and the least affected by a terrorist attack.

## Attacks Over Time by Location

Another visualization we utilized to see how many deaths occurred in a single year in each country overtime was an animated choropleth map. This analysis required a bit of grouping in pandas to calculate different sums for each nation. First we grouped by the year, then country, and summed the number killed. Then, we reset the index to use the sub-dataframe in an animation and set each of the variables in the plotly express code. With this map, the time frame, or year, is seen at the bottom as the scroller plays through the past 2 decades and the number of people killed in a country in that given year is indicated on the legend to the right.

This map also provides some insight on how many people are killed in terrorist attacks in a given year over time across the globe. Here, we see that although 9/11 was the largest attack in this dataset, other countries have many more attacks in a given year that kill more than that single attack. The legend ranges to 12,000 deaths and countries with those deaths in a given year tend to be in the East and smaller in size. For example, Iraq and Afghanistan are 13 of the top 15 country deaths per year, and they are smaller nations in size. This is particularly noticeable in the last decade, where these countries pop out with colors of green, yellow, orange, and red, indicating that 8,000-12,000 deaths occurred in those years.

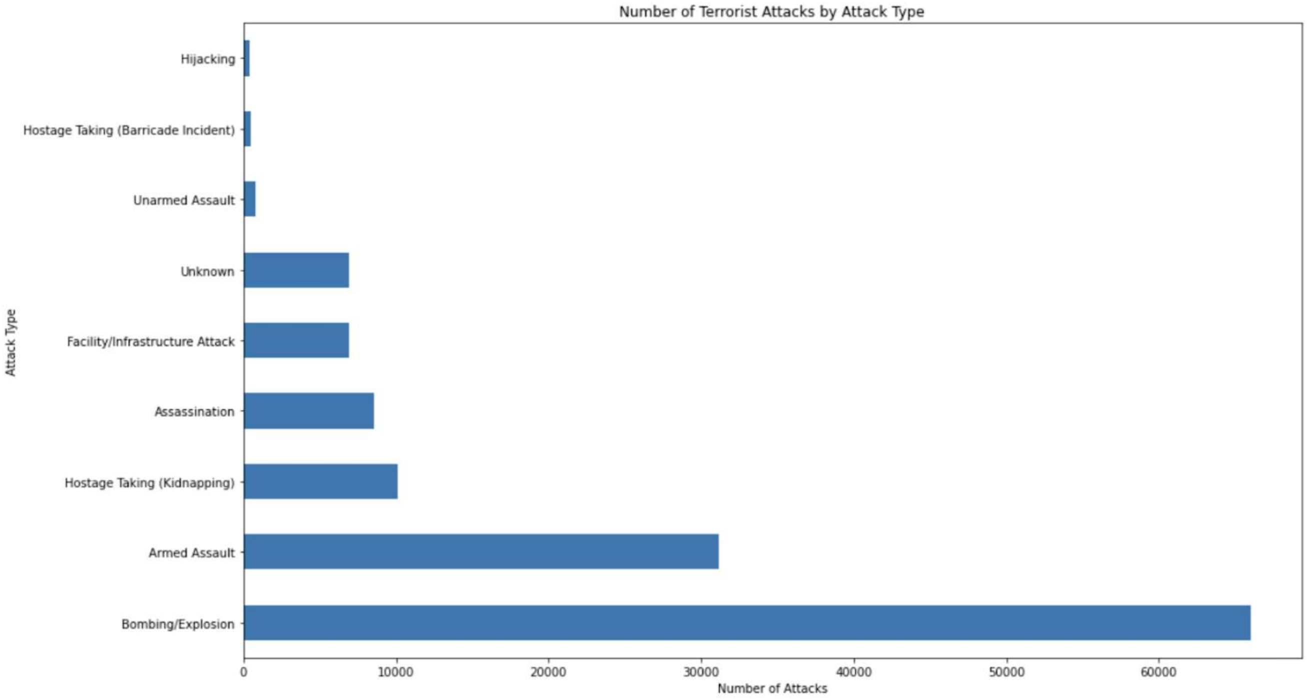


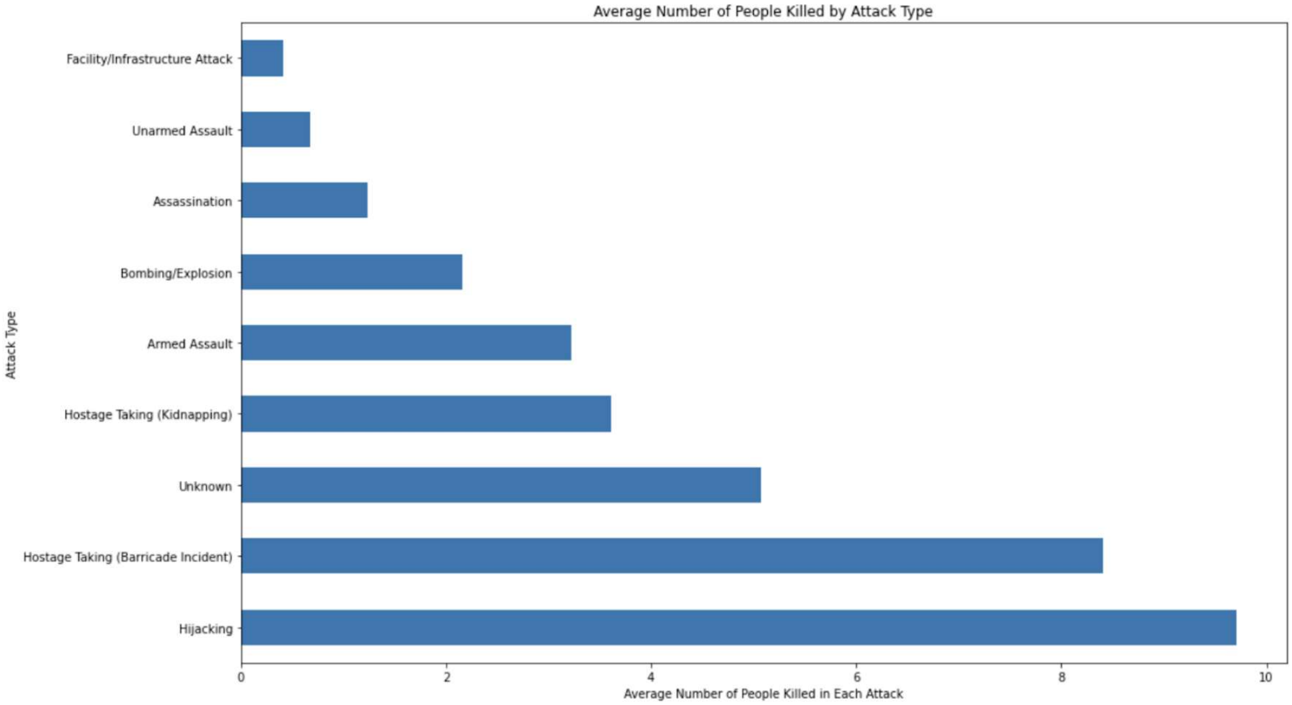
## Motive Analysis

Arguably, the most important question one would ask themselves when studying data of this nature is why would anyone commit terror attacks? When we looked at the frequency of words and the word cloud for motives, we saw that oftentimes motives are unknown. This analysis was done with the NLTK library as well as the WordCloud library. With these packages, we noted that the term ‘however’ is the 5th most frequent word. This could open the door for further text exploratory analysis to study what circumstances lead researchers to assume certain groups committed attacks. Additionally, we could look into why some groups openly claim responsibility while others do not. We could narrow down the motive categories into political, economic, and religious, and try to analyze if there are any differences between these open claims of responsibility versus those groups that do not.

## Attack Type Analysis

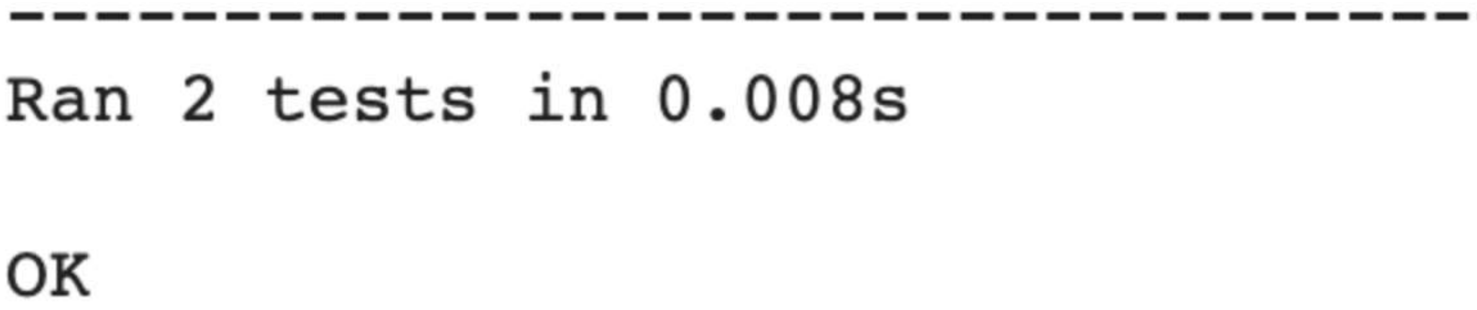
When it comes to how a terrorist attack occurs, there are many different ways. In this dataset there were eight different attack types plus a category for unknown attack type. The first chart is the number of unique attacks by attack type. We can see that bombings and armed assaults were the most prominent types. The second chart is the average number of people killed in each attack type. We can see that hijacking is the deadliest attack type, killing an average of almost 10 people per attack. It is probable that this value is being inflated by the events of 9/11 which is one of the deadliest terrorist attacks in this dataset.





## Testing

Because we focused our analysis on the most recent two decades, we had to first confirm that the code correctly subsetted the entire dataset. We did this but first subsetting the data to remove all the years prior to 2000. Afterwards, we confirmed that the number of rows for years prior to 2000 was zero. This test passed. We also tested our code that grouped the number of terrorist attacks each month for every year in the dataset. We did this by calling the groupby function on year and month, counting the number of rows in the groupby object then confirming that the resulting number of rows was equal to the number we expected. This test passed as well.



```
---------------------------------
Ran 2 tests in 0.008s


OK
```

## Explanation of Results & Conclusions

From a time perspective, we can see grouped events on a micro and macro scale. From a location perspective we can see hotspots on earth where there is more activity. From a motive perspective we can see the impact of events on fatalities. All three of these can help point us towards patterns to eventually create a model to predict and prevent future attacks.

To take this research forward, we would use our aggregated datapoints to feed into a machine learning model. This could help us identify the more detailed connections between all three subjects.

*Photo from https://www.history.com/topics/21st-century/9-11-timeline*