Final Project Step 2

Stewart Wilson

2022-05-22

How to Import and Clean My Data

The first step is to import and clean the data. I import the data from the terrorism database, military spending, and troop deployment datasets into their own data frame. I then use the head(), tail(), and names() functions to get an idea of what the data looks like.

Let's look at each database individually.

Terrorism Databse

```
##   [1] "eventid"           "iyear"             "imonth"
##   [4] "iday"              "approxdate"        "extended"
##   [7] "resolution"        "country"           "country_txt"
##  [10] "region"            "region_txt"        "provstate"
##  [13] "city"              "latitude"          "longitude"
##  [16] "specificity"       "vicinity"          "location"
##  [19] "summary"           "crit1"             "crit2"
##  [22] "crit3"             "doubtterr"         "alternative"
##  [25] "alternative_txt"   "multiple"          "success"
##  [28] "suicide"           "attacktype1"       "attacktype1_txt"
##  [31] "attacktype2"       "attacktype2_txt"   "attacktype3"
##  [34] "attacktype3_txt"   "targtype1"         "targtype1_txt"
##  [37] "targsubtype1"      "targsubtype1_txt"  "corp1"
##  [40] "target1"           "natlty1"           "natlty1_txt"
##  [43] "targtype2"         "targtype2_txt"     "targsubtype2"
##  [46] "targsubtype2_txt"  "corp2"             "target2"
##  [49] "natlty2"           "natlty2_txt"       "targtype3"
##  [52] "targtype3_txt"     "targsubtype3"      "targsubtype3_txt"
##  [55] "corp3"             "target3"           "natlty3"
##  [58] "natlty3_txt"       "gname"             "gsubname"
##  [61] "gname2"            "gsubname2"         "gname3"
##  [64] "gsubname3"         "motive"            "guncertain1"
##  [67] "guncertain2"       "guncertain3"       "individual"
##  [70] "nperps"            "nperpcap"          "claimed"
##  [73] "claimmode"         "claimmode_txt"     "claim2"
##  [76] "claimmode2"        "claimmode2_txt"    "claim3"
##  [79] "claimmode3"        "claimmode3_txt"    "compclaim"
##  [82] "weaptype1"         "weaptype1_txt"     "weapsubtype1"
##  [85] "weapsubtype1_txt"  "weaptype2"         "weaptype2_txt"
##  [88] "weapsubtype2"      "weapsubtype2_txt"  "weaptype3"
##  [91] "weaptype3_txt"     "weapsubtype3"      "weapsubtype3_txt"
```

```
##  [94] "weaptype4"          "weaptype4_txt"      "weapsubtype4"
##  [97] "weapsubtype4_txt"   "weapdetail"         "nkill"
## [100] "nkillus"            "nkillter"           "nwound"
## [103] "nwoundus"           "nwoundte"           "property"
## [106] "propextent"         "propextent_txt"     "propvalue"
## [109] "propcomment"        "ishostkid"          "nhostkid"
## [112] "nhostkidus"         "nhours"             "ndays"
## [115] "divert"             "kidhijcountry"      "ransom"
## [118] "ransomamt"          "ransomamtus"        "ransompaid"
## [121] "ransompaidus"       "ransomnote"         "hostkidoutcome"
## [124] "hostkidoutcome_txt" "nreleased"          "addnotes"
## [127] "scite1"             "scite2"             "scite3"
## [130] "dbsource"           "INT_LOG"            "INT_IDEO"
## [133] "INT_MISC"           "INT_ANY"            "related"
```

There are over a 100 variables used in the terrorism database; I need to filter that down towards a manageable numbner for the sake of brevity and computer memory when I start combining the datasets. I choose the variables

- iyear
- imonth
- country_txt
- region
- success
- nkill

as they seem the most informative for the purposes of viewing terrorism trends broadly.

Troop Deployment Database

```
##          country code iso3c year troops army navy   air marine
## 1 United Kingdom  200   GBR 2006  11331  397  584 10280     70
## 2 United Kingdom  200   GBR 2007  10425  355  443  9552     75
## 3 United Kingdom  200   GBR 2008   9042  315  489  8169     69
## 4 United Kingdom  200   GBR 2009   8933  324  396  8143     70
## 5 United Kingdom  200   GBR 2010   8764  333  364  8004     63
## 6 United Kingdom  200   GBR 2011   8673  328  316  7977     52
```

```
##     country code iso3c year troops army navy air marine
## 274 Denmark  390   DNK 2010     10    2    4   4      0
## 275 Denmark  390   DNK 2011     10    2    4   4      0
## 276 Denmark  390   DNK 2012     19    3    5   5      6
## 277 Denmark  390   DNK 2013     16    2    4   5      5
## 278 Denmark  390   DNK 2014     14    2    4   4      4
## 279 Denmark  390   DNK 2015     15    2    3   4      6
```

```
## [1] "country" "code"    "iso3c"   "year"    "troops"  "army"    "navy"
## [8] "air"     "marine"
```

The troop deployment database, only covers troop deployment in the EU countries, so it does not include troop deployment to Afghanistan or Iraq, our key areas of interest. For now, I leave it be, and I will return to it if I need to get more detailed about terrorist attacks in Europe specifically.

Military Spending Dataset

```
##   Year DefenseBudget   GDP Population
## 1 1960         47.35 543.3     180.67
## 2 1961         49.88 563.3     183.69
## 3 1962         54.65 605.1     186.54
## 4 1963         54.56 638.6     189.24
## 5 1964         53.43 685.8     191.89
## 6 1965         54.56 743.7     194.30
```

```
##    Year DefenseBudget      GDP Population
## 56 2015        633.83 18238.30     320.64
## 57 2016        639.86 18745.08     322.94
## 58 2017        646.75 19542.98     324.99
## 59 2018        682.49 20611.86     326.69
## 60 2019        731.75 21433.22     328.24
## 61 2020        778.00 20940.00     330.66
```

```
## [1] "Year"          "DefenseBudget" "GDP"           "Population"
```

This dataset looks good. However, I will need to be careful when I merge it to make sure the years match with the years of terrorist attacks.

Merging Datasetes and Final Cleaning

Now I merge the terrorist database with the military spending database. I do a full_join() in order to keep all columns from the two datasets.

Now that the two datasets are one, I use rename_with() to make all the column titles the same case for ease of later analysis. I also remove datapoints from before 1970 since the terrorism database starts in 1970

A Look at the Final Data Set

And here is the final data set we will be using (sliced to show three different parts of the dataframe):

```
##    iyear imonth        country_txt region success nkill defensebudget      gdp
## 1   1970      7 Dominican Republic      2       1     1         83.41  1073.30
## 2   1970      0              Mexico      1       1     0         83.41  1073.30
## 3   1970      1         Philippines      5       1     1         83.41  1073.30
## 4   1970      1              Greece      8       1    NA         83.41  1073.30
## 5   1970      1               Japan      4       1    NA         83.41  1073.30
## 6   1992      1             Germany      8       1     3        325.03  6520.33
## 7   1992      1      United Kingdom      8       1     0        325.03  6520.33
## 8   1992      1      United Kingdom      8       1     0        325.03  6520.33
## 9   1992      1      United Kingdom      8       1     0        325.03  6520.33
## 10  1992      1         Philippines      5       1     0        325.03  6520.33
## 11  1992      1              Panama      2       1     0        325.03  6520.33
## 12  2017      5                Iraq     10       0     0        646.75 19542.98
## 13  2017      5                Iraq     10       0     0        646.75 19542.98
## 14  2017      5                Iraq     10       0     1        646.75 19542.98
## 15  2017      5                Iraq     10       0     0        646.75 19542.98
```

```
## 16  2017      5              Iraq     10       1     0        646.75 19542.98
## 17  2017      5              Iraq     10       1    12        646.75 19542.98
##    population
## 1      205.05
## 2      205.05
## 3      205.05
## 4      205.05
## 5      205.05
## 6      256.51
## 7      256.51
## 8      256.51
## 9      256.51
## 10     256.51
## 11     256.51
## 12     324.99
## 13     324.99
## 14     324.99
## 15     324.99
## 16     324.99
## 17     324.99
```

Future Step Questions

One thing that I need to continue learning how to do, is investigating details in the dataframe that I may not see through just looking at the head and tail but could impact my analysis. For instance, how can I check for misspellings of country names?

In general, I should investigate if, in the case of my analysis particularly, it is a better idea to ignore nans and if not, how best to handle them.

The biggest issue is since the data prior to the 2000s was collected retroactively, how do I account for the difference in data collection? Or rather, how might the difference in how data was collected during some years affect my analysis?

Perhaps most importantly, I need to continue looking for a dataset on troop deployment in Afghanistan and Iraq.

What Information Is Not Self-Evident?

Information that is not readily available includes:

- Have terrorist attacks increased or decreased in the last 50 years?
- What has the impact of increased military spending been on terrorism rates?
- Where are terrorist attacks most focused?

How Could I Look at the Data

I could look at the Data in a number of ways. I will filter by country, year, and defense budget. I could create a new data set that is just an accumulation of number of terrorist attacks per year, which may help simplify the work.

I should split the data set into before 9/11 and after 9/11 and observe the differences. This will provide me insight into the effects the War on Terror has had.

I will also want to do a correlation analysis between terrorist attacks and military spending.

Manipulation Plan

In sum, my plan is to slice the data set into a number of subsets and then I will compare those data sets to one another to look for trends and differences.The manipulations I plan to do are as follows:

- Split dataset into before 9/11 and after 9/11
- Arrange by number killed
- Create data set for US terrorist attacks
- Correlation analysis of terrorist attacks

Data Summary

After manipulating the data following the plan above, we can summarize the data initially as follows:

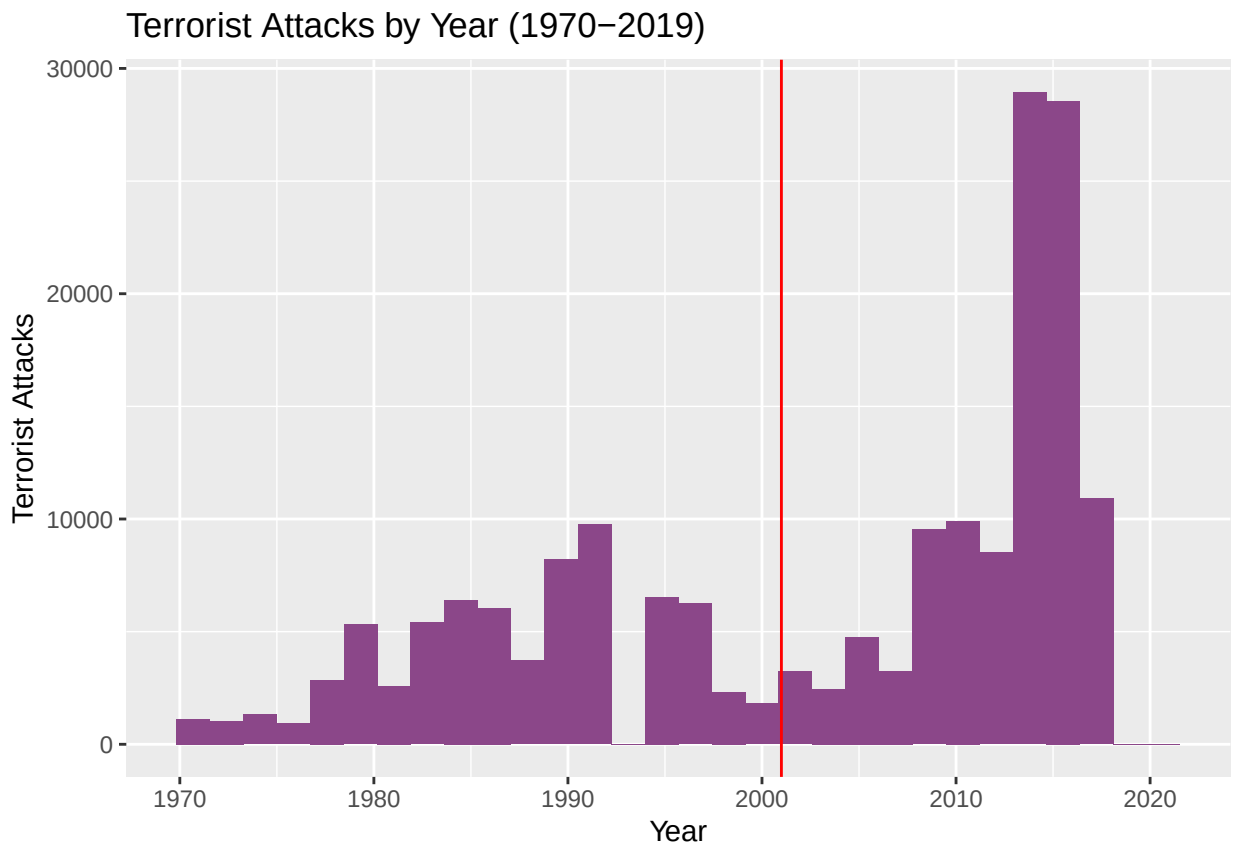
- There were 71,661 terrorist attacks worldwide recorded between 1970 and 2001
- There were 110,044 terrorist attacks worldwide recorded between 2001 and 2019
- There were 2,434 terrorist attacks in the US recorded between 1970 and 2001
- There were 402 terrorist attacks in the US recorded between 2001 and 2019
- The total military expenditure was $7,779.56 (in billions) between 1970 and 2001
- The total military expenditure was $12,011.01 (in billions) between 2001 and 2020
- The mean military expenditure was $185.23 (in billions) between 1970 and 2001
- The mean military expenditure was $632.15 (in billions) between 2001 and 2020
- There is a .449 correlation coefficient between attacks per year and military spending per year
  - What is interesting here is that there is a positive correlation, meaning the more military expenditure there is, the more terrorist attacks there are
- Military spending accounts for 20.2% of the variability in terrorist attacks
- The p-value between terrorist attacks per year and military spending per year is .00092
  - This means it is unlikely the relationship is due by chance

Plots and Tables

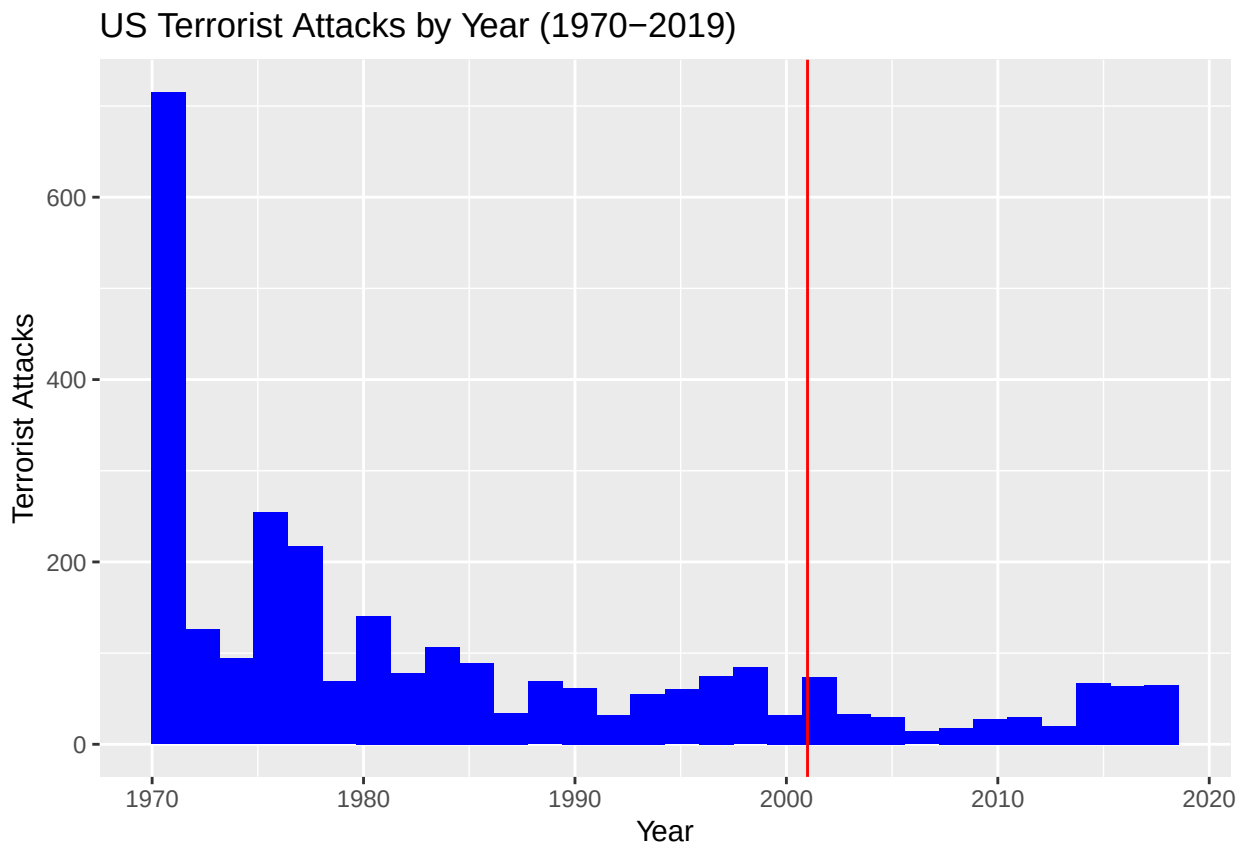
The above findings are pretty hideous to read and understanding what they mean can be difficult. We can better visualize these findings through histograms, bar graphs, scatterplots, and tables.

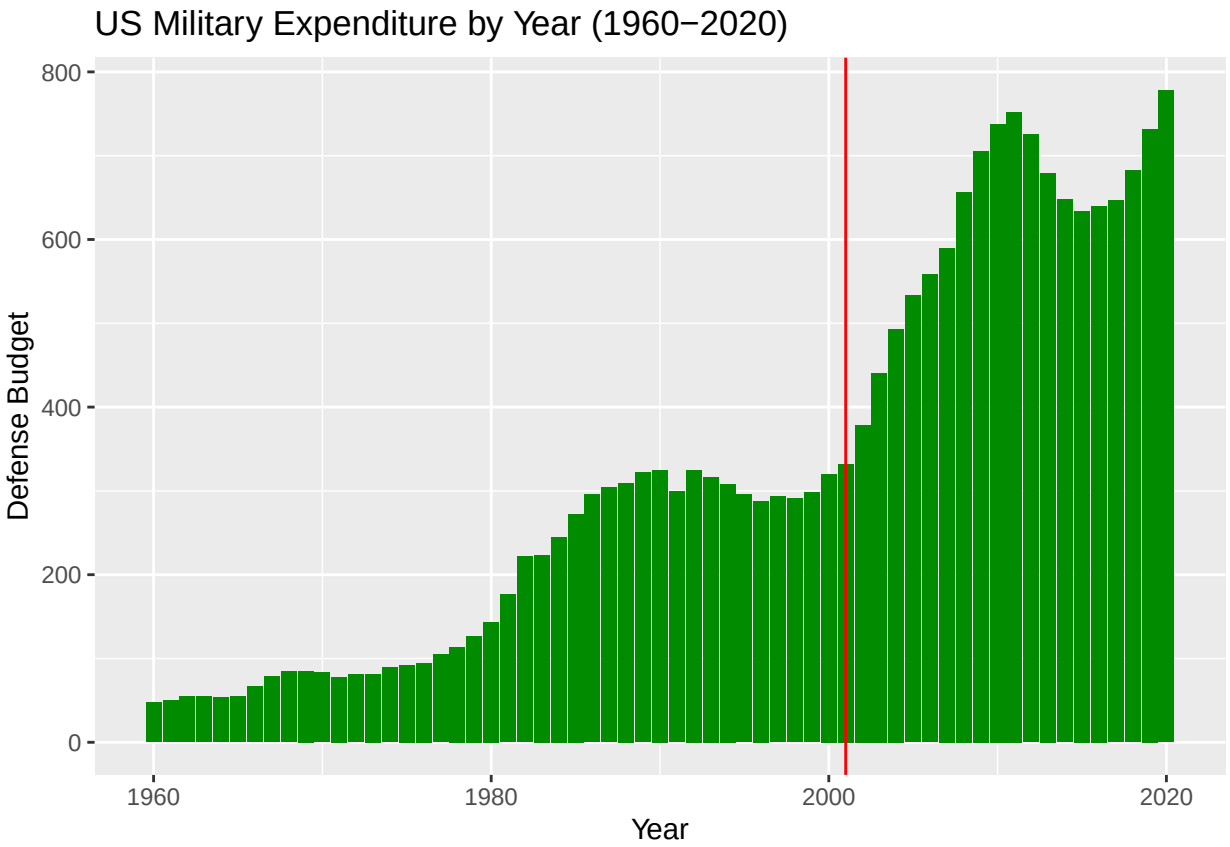
NOTE: a vertical red line is added to mark the year of 9/11

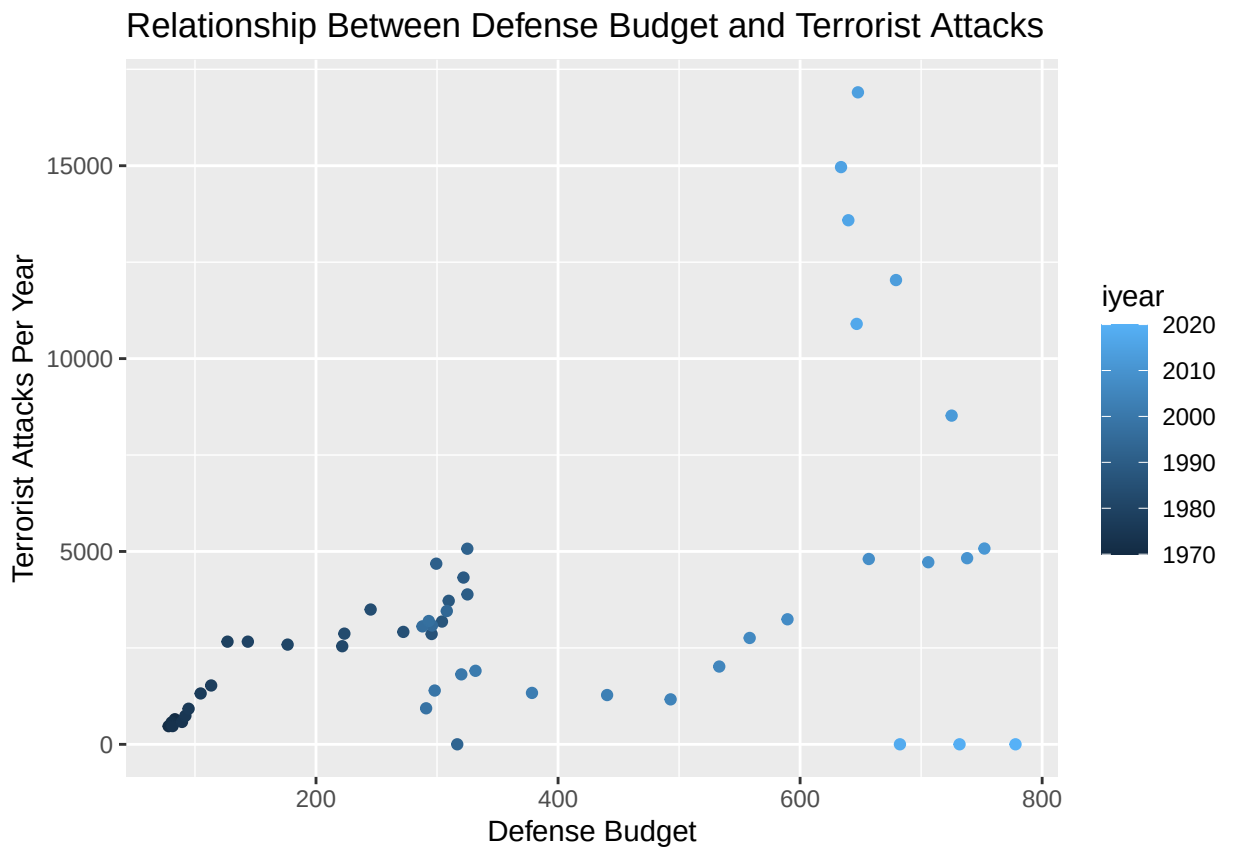
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


US Terrorist Attacks by Year (1970−2019)
Terrorist Attacks
600
400
200
0
1970
1980
1990
2000
2010
2020
Year


US Military Expenditure by Year (1960−2020)
Defense Budget
800
600
400
200
0
1960
1980
2000
2020
Year



To help put the correlation coefficient between Defense Budget and Terrorist Attacks into perspective, I did correlation tests using US GDP and population as well. The table below summarizes those findings.

	Year	Attacks	DefenseBudget	GDP	Population
Year	1.0000000	0.3900066	0.9494118	0.9996380	1.0000000
Attacks	0.3900066	1.0000000	0.4499763	0.3894635	0.3900066
DefenseBudget	0.9494118	0.4499763	1.0000000	0.9481448	0.9494118
GDP	0.9996380	0.3894635	0.9481448	1.0000000	0.9996380
Population	1.0000000	0.3900066	0.9494118	0.9996380	1.0000000

Machine Learning?

I plan on doing some regression analysis to further investigate the relationship between defense spending and terrorism rates worldwide.

In addition, I plan on converting the datasets I have into time series and doing time series analysis to have a more accurate reading of the changes over time to my data. # Future Steps I have several steps ahead of me as I continue my . As I said above, doing regression analysis and time series analysis are my next steps. I also need to look over my code to look for ways to make it more efficient.

Further, the distribution of the data does not follow a normal distribution. I should investigate if transforming some of the variables will create a normal distribution.

I will also continue searching for datasets on US involvement in Afghanistan and Iraq. While defense spending provides some glimpse into US military action, it is a very general look that would be best supported by data dealing with troop deployment.

That said, the biggest thing ahead of me is analyzing my assumptions as I move forward. I quickly realized during this step of the analysis that my research questions were far too broad to be answered simply in a single analysis. Moreover, there was data I needed to conduct my analysis that were simply not available or I have yet to find them. As such, I needed to narrow in on the data I had to see what questions I *could* actually respond to. This has led me to investigating military spending and attacks per year specifically. Yet, this comes with a series of assumptions I need to investigate in order to ensure my conclusions are significant and meaningful.