EDA for armed conflict data

Maksim Helmann

2024-09-30

Load data

Before proceeding with the EDA, we first load the data using the `read.csv()` function. The dataframe contains the following column names:

```
 [1] "country_name" "ISO"          "region"       "year"         "gdp1000"
 [6] "OECD"         "OECD2023"     "popdens"      "urban"        "agedep"
[11] "male_edu"     "temp"         "rainfall1000" "totdeath"     "armconf1"
[16] "matmor"       "infmor"       "neomor"       "un5mor"       "drought"
[21] "earthquake"
```

And the first few rows:

```
  country_name ISO         region year   gdp1000 OECD OECD2023  popdens    urban
1  Afghanistan AFG Southern Asia 2000        NA    0        0 14.13654 16.25324
2  Afghanistan AFG Southern Asia 2001        NA    0        0 14.23156 16.25661
3  Afghanistan AFG Southern Asia 2002 0.1835328    0        0 14.32270 16.42654
4  Afghanistan AFG Southern Asia 2003 0.2004626    0        0 14.40691 16.60701
5  Afghanistan AFG Southern Asia 2004 0.2216576    0        0 15.21947 16.71367
6  Afghanistan AFG Southern Asia 2005 0.2550551    0        0 15.33619 16.85096
    agedep male_edu     temp rainfall1000 totdeath armconf1 matmor infmor
1 108.3466 2.762086 12.69959    0.2763704     5065        1   1450   90.5
2 108.9899 2.856936 12.85570    0.2793079     5394        1   1390   87.9
3 109.3472 2.954241 12.71081    0.3805710     5553        1   1300   85.3
4 109.4475 3.054121 12.16592    0.4288939     1157        1   1240   82.7
5 109.2868 3.156706 13.04643    0.3754336      944        1   1180   80.0
6 107.9646 3.262133 12.23141    0.4415680      817        1   1140   77.3
  neomor un5mor drought earthquake
1   60.9  129.2       0          1
2   59.7  125.2       1          0
```

```
3   58.5  121.1       1          0
4   57.2  116.9       1          0
5   55.9  112.6       1          0
6   54.6  108.4       1          0
```

Additionally, we want to take a look at the random selection to eventually observe any anomalies:

```
        country_name ISO             region year    gdp1000 OECD OECD2023
1       Sierra Leone SLE Sub-Saharan Africa 2013  0.7064527    0        0
2       Sierra Leone SLE Sub-Saharan Africa 2014  0.7023354    0        0
3 Russian Federation RUS     Eastern Europe 2002  2.3775295    0        0
4            Vanuatu VUT          Melanesia 2019  3.0765899    0        0
5         Bangladesh BGD      Southern Asia 2019  2.1220789    0        0
6             Kuwait KWT       Western Asia 2019 30.6673482    0        0
   popdens      urban   agedep  male_edu      temp rainfall1000 totdeath
1 24.20579 22.3032083 84.20602  3.977973 26.586898   2.58806965        0
2 24.28193 22.5946809 83.00946  4.088727 26.535845   2.58356012        0
3 41.81448 31.1748659 41.71371 11.913114  5.005679   0.57717724      992
4 23.94911  0.6392962 77.10057  7.058168 24.118731   2.51996230        0
5 77.72937 37.9023785 49.25528  5.928060 25.477084   2.72341782       28
6 70.10013 65.1536080 32.20743 11.037412 27.174540   0.09664143        0
  armconf1 matmor infmor neomor un5mor drought earthquake
1        0   1180   97.7   36.5  141.2       0          0
2        0   1190   95.6   35.7  139.2       0          0
3        1     51   13.8    8.0   17.2       0          0
4        0     NA   21.6   10.9   25.6       0          0
5        1     NA   25.5   18.5   30.7       0          0
6        0     NA    7.5    4.9    8.8       0          0
```

Key summary statistics

Next, we are interested in understanding key summary statistics, such as the minimum, median, and maximum values for numeric and binary variables with summary() from base R and the number of observations.

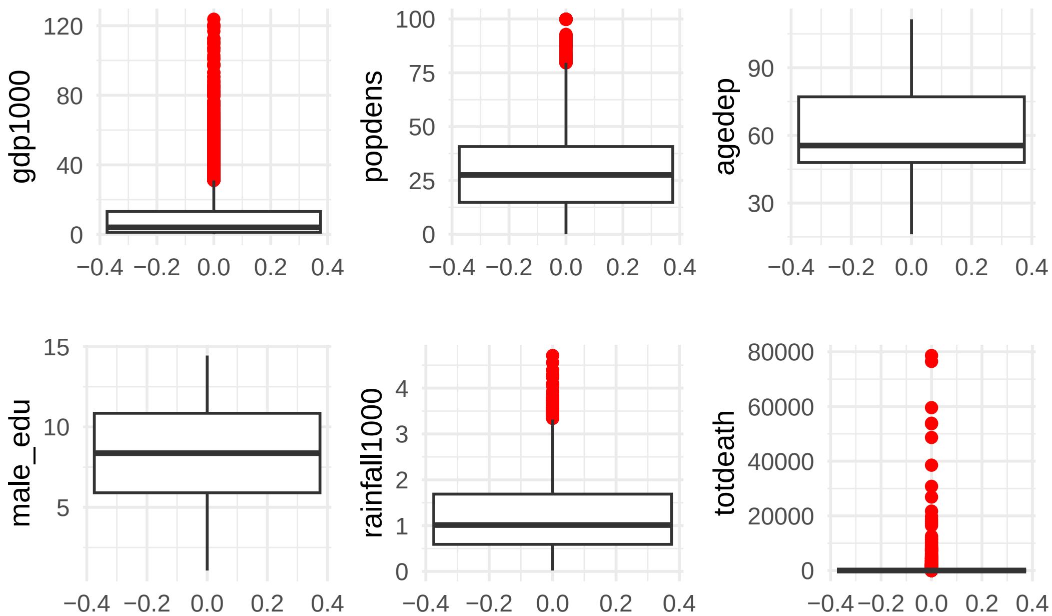
```
      year         gdp1000              OECD           OECD2023
 Min.   :2000   Min.   :  0.1105   Min.   :0.000   Min.   :0.0000
 1st Qu.:2005   1st Qu.:  1.2383   1st Qu.:0.000   1st Qu.:0.0000
 Median :2010   Median :  4.0719   Median :0.000   Median :0.0000
 Mean   :2010   Mean   : 11.4917   Mean   :0.171   Mean   :0.1882
```

```
 3rd Qu.:2014   3rd Qu.: 13.1531   3rd Qu.:0.000   3rd Qu.:0.0000  
 Max.   :2019   Max.   :123.6787   Max.   :1.000   Max.   :1.0000  
                NA's   :62                                         
    popdens          urban              agedep          male_edu     
 Min.   : 0.00   Min.   : 0.1025   Min.   : 16.17   Min.   : 1.067  
 1st Qu.:14.79   1st Qu.:17.2872   1st Qu.: 47.94   1st Qu.: 5.904  
 Median :27.52   Median :30.2535   Median : 55.51   Median : 8.368  
 Mean   :30.57   Mean   :30.6948   Mean   : 61.94   Mean   : 8.258  
 3rd Qu.:40.72   3rd Qu.:41.6558   3rd Qu.: 77.11   3rd Qu.:10.849  
 Max.   :99.86   Max.   :93.4135   Max.   :111.48   Max.   :14.441  
 NA's   :20      NA's   :20                         NA's   :20      
      temp         rainfall1000        totdeath          armconf1     
 Min.   :-2.405   Min.   :0.01993   Min.   :    0.0   Min.   :0.0000  
 1st Qu.:12.928   1st Qu.:0.59146   1st Qu.:    0.0   1st Qu.:0.0000  
 Median :21.958   Median :1.01288   Median :    0.0   Median :0.0000  
 Mean   :19.625   Mean   :1.20216   Mean   :  361.1   Mean   :0.1892  
 3rd Qu.:25.869   3rd Qu.:1.68706   3rd Qu.:    2.0   3rd Qu.:0.0000  
 Max.   :29.676   Max.   :4.71081   Max.   :78644.0   Max.   :1.0000  
 NA's   :20       NA's   :20                                          
     matmor           infmor            neomor          un5mor      
 Min.   :   2.0   Min.   :  1.60   Min.   : 0.80   Min.   :  2.00  
 1st Qu.:  17.0   1st Qu.:  7.60   1st Qu.: 4.90   1st Qu.:  9.00  
 Median :  66.0   Median : 18.90   Median :12.10   Median : 22.20  
 Mean   : 210.6   Mean   : 28.90   Mean   :16.18   Mean   : 40.50  
 3rd Qu.: 299.8   3rd Qu.: 44.52   3rd Qu.:25.32   3rd Qu.: 61.33  
 Max.   :2480.0   Max.   :138.10   Max.   :60.90   Max.   :224.90  
 NA's   :426      NA's   :20       NA's   :20      NA's   :20      
    drought          earthquake     
 Min.   :0.00000   Min.   :0.00000  
 1st Qu.:0.00000   1st Qu.:0.00000  
 Median :0.00000   Median :0.00000  
 Mean   :0.08333   Mean   :0.08737  
 3rd Qu.:0.00000   3rd Qu.:0.00000  
 Max.   :1.00000   Max.   :1.00000  
```

Looking at the summary statistic, we can note a few key observations. First, there are a few covariates with missing values like GDP (`gdp1000`, 62 missing) and maternal mortality (`matmor`, 426 missing). Second, wide ranges can be observed in GDP (0.11 to 123.68), suggesting large economic disparities, and total deaths (up to 78,644), likely reflecting countries or events with high mortality. High infant mortality (up to 138.10) and a wide variation in under-5 mortality (2 to 224.9) point to severe health challenges in certain regions.

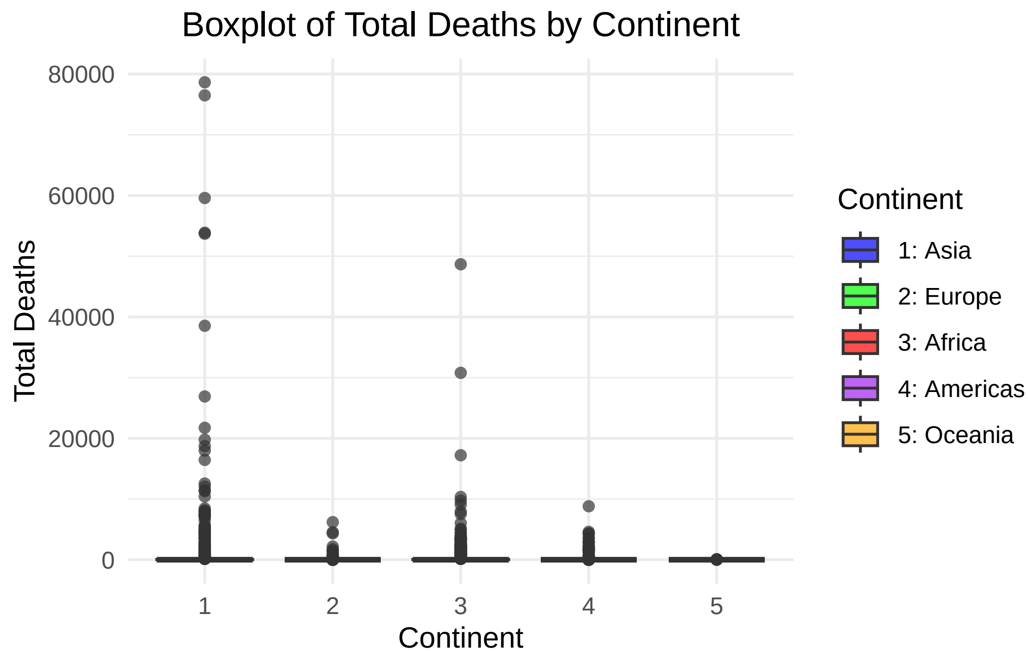
Identifying Skewness and Symmetry

In the following we want to look at some boxplots to examine the summary statistics visiually and to detect patterns or anomalies in the data.

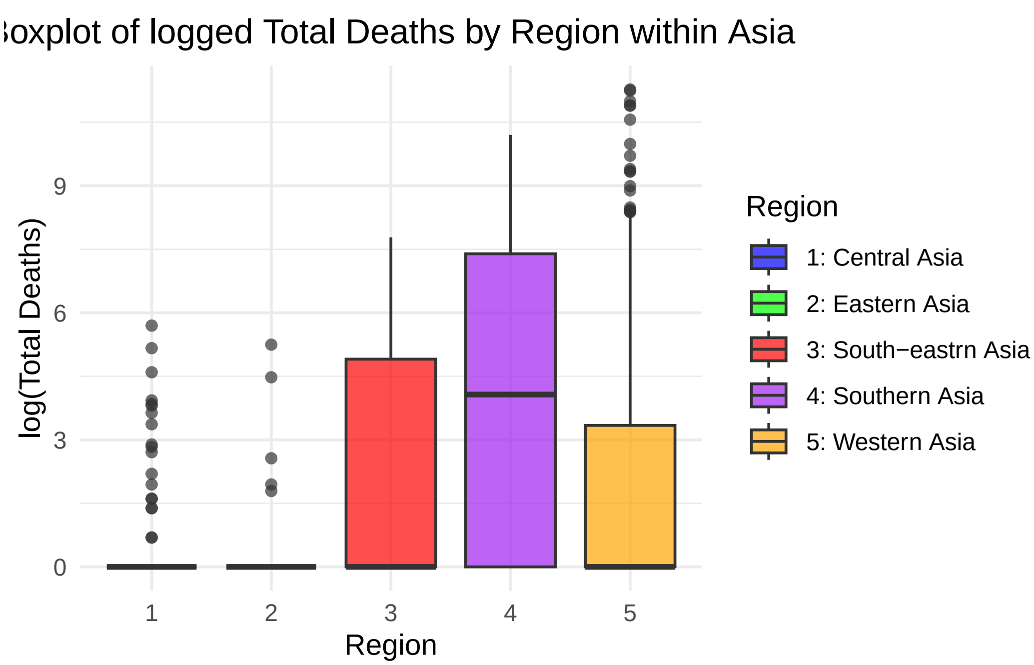


The examined variables are GDP per 1000 (`gdp1000`), population density (`popdens`), age dependency ratio (`agedep`), male education (`male_edu`), rainfall per 1000 mm (`rainfall1000`), and total deaths (`totdeath`). For most variables, the distribution shows some level of skewness with significant outliers, especially in gdp1000 and totdeath. These variables have outliers that extend far beyond the interquartile range, suggesting a number of extreme values or observations. On the other hand, variables like male_edu and agedep show relatively more symmetric distributions, with fewer outliers or extreme values. Therefore, it would be reasonable to look at the logged values of GDP to get a better sense of the distribution. For the totdeath variable we first inspect the boxplot for each "continent" because visual inspection using a histogram did not result in meaningul observations as the outliers really skew the plot. The grouping is achieved by grouping corresponding regions into one group. Afterward, we will decide whether we take a log transform or consider the variable to be binary or categorical.

The corresponding groups for the regions are the following: 1) Asia: Southern Asia, Western Asia, Eastern Asia, South-eastern Asia, Central Asia 2) Europe: Southern Europe, Western Europe, Eastern Europe, Northern Europe 3) Africa: Northern Africa, Sub-Saharan Africa 4) Americas: Northern America, Latin America and the Caribbean 5) Oceania: Australia and New Zealand, Micronesia, Melanesia, Polynesia



For the group Asia and Africa we observe a high level of skewness with significant outliers. Next, we want to find out which region is the most skewed. After that, we will compute the proportion of armed conflicts in those two groups.



Before taking the log of the total death variable, we inspected the boxplots for the original scale. The boxplots showed that the data for total deaths is heavily skewed, with several extreme outliers, particularly in Southern Asia and Western Asia. The scale is dominated by these outliers, as the majority of the data points are clustered near zero, making the overall distribution hard to interpret clearly. In addition, this indicates that these regions experienced significantly higher death tolls compared to the others. On the other hand, Central and Eastern Asia, regions show relatively minimal variation, with their total deaths clustered near zero, and they have few or no outliers, suggesting that in these regions, the total deaths tend to be consistently low.

After applying the logarithmic transformation to the total deaths, the spread of the data across the regions becomes much clearer. The log transformation reduces the influence of extreme outliers and compresses the range of high values, making it easier to compare the regions on a more consistent scale. All regions now show some degree of spread in the data, especially South-eastern Asia, Southern Asia, and Western Asia, indicating a more balanced distribution of log-transformed total deaths across these regions. Even with the log transformation, Southern Asia and Western Asia still exhibit several outliers, indicating that there are still extreme cases of high total deaths, though they are less extreme compared to the first figure without the log transformation.

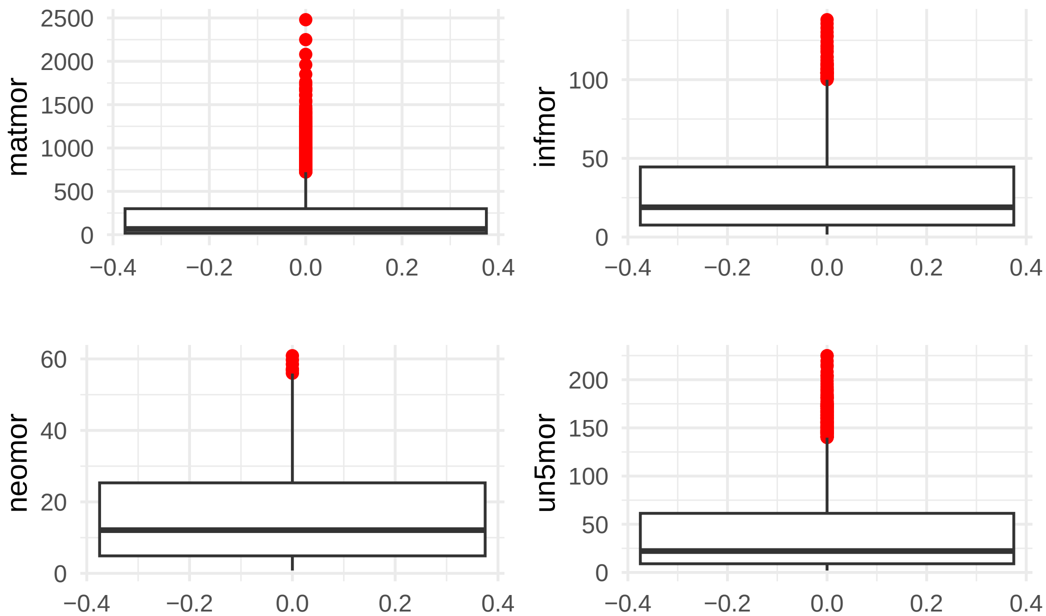
Below the proportion of armed conflicts in different regions is depicted in a table:

```
# A tibble: 17 x 2
   region                            prop
   <chr>                            <dbl>
 1 Australia and New Zealand       0
 2 Central Asia                    0.08
 3 Eastern Asia                    0.02
 4 Eastern Europe                  0.125
 5 Latin America and the Caribbean 0.127
 6 Melanesia                       0.0375
 7 Micronesia                      0
 8 Northern Africa                 0.533
 9 Northern America                0.025
10 Northern Europe                 0.005
11 Polynesia                       0
12 South-eastern Asia              0.305
13 Southern Asia                   0.561
14 Southern Europe                 0.0115
15 Sub-Saharan Africa              0.269
16 Western Asia                    0.262
17 Western Europe                  0.0214
```

Among all the regions, Southern Asia has the highest proportion of armed conflicts with a value of 0.5611 indicating that over half of the occurrences are in this region. Northern

Africa has the second-highest proportion at 0.5333. South-eastern Asia and Sub-Saharan Africa also show relatively high proportions of conflicts at 0.3045 (30.5%) and 0.2688 (26.9%) respectively. Northern America, Western Europe, Eastern Asia, and Northern Europe all have low proportions of conflicts, with values ranging from 0.0050 to 0.0250.

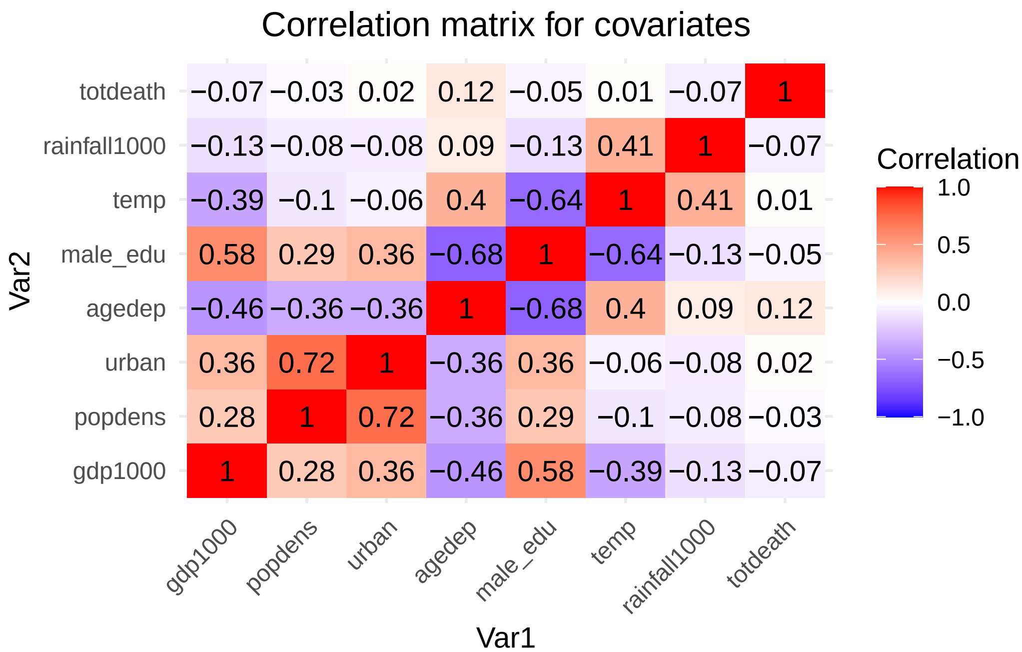
Below the boxplots for the outcome variables are provided.



From the boxplots we can observe that all mortality variables have some extreme outliers, especially the maternal mortality. In addition, `matmor` has 426 missing values, while the rest has each 20 missing values. Furthermore, most of the data for all four variables is concentrated near zero, indicating that for the majority of the observations, mortality rates are relatively low. However, the presence of significant outliers suggests that certain regions or cases experience much higher mortality rates, especially for maternal mortality. The distribution of values in each boxplot appears to be positively skewed, with the majority of data points lying close to zero and a few extreme values extending the whiskers upward.

Correlaiton withing region

We first want to identify how many unique regions the dataframe contains:



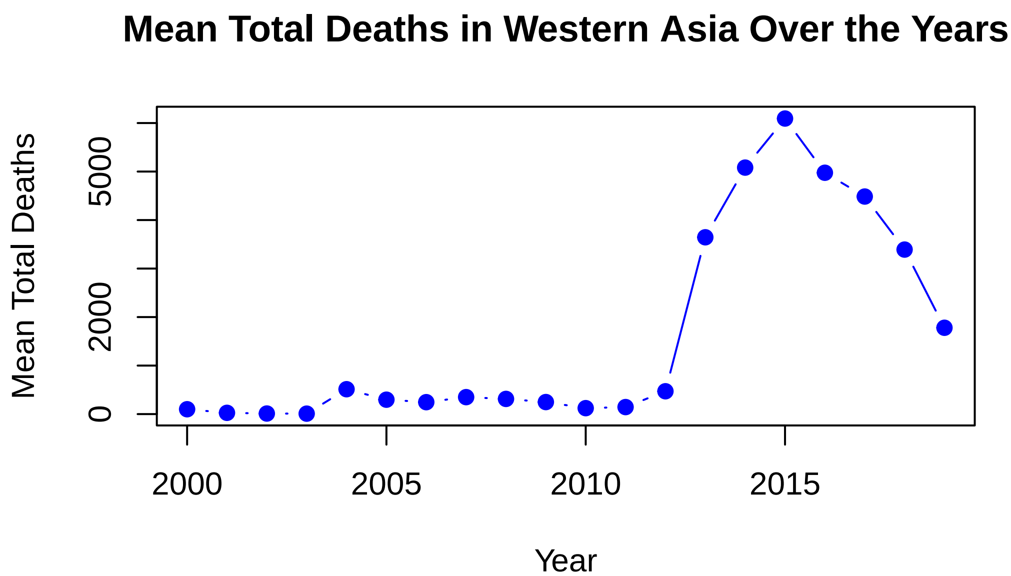
The correlation matrix reveals several key relationships among the variables. Notably, `male_edu` shows a strong positive correlation with `gdp1000` (0.58), indicating that higher male education levels are associated with higher GDP per capita. There is also a significant positive correlation between `urban` and `popdens` (0.72), suggesting that higher population density correlates with greater urbanization. Conversely, `male_ede` is strongly negatively correlated with `agedep` (-0.68), implying that as the male education level increases, age dependency decreases. Additionally, `temp` negatively correlates with both `gdp1000` (-0.46) and `male_edu` (-0.64), indicating that higher temperatures are associated with lower GDP and male education levels.

Data visualization for outcome variables

```
          matmor    infmor    neomor    un5mor
matmor 1.0000000 0.8785612 0.8354908 0.8994877
infmor 0.8785612 1.0000000 0.9590878 0.9861117
neomor 0.8354908 0.9590878 1.0000000 0.9278720
un5mor 0.8994877 0.9861117 0.9278720 1.0000000
```

The correlation matrix reveals strong positive correlations among the four variables: `matmor`, `infmor`, `neomor`, and `un5mor`. The correlations suggest that increases in one mortality type are associated with increases in the others, highlighting potential underlying factors contributing to overall mortality rates in certain regions. The strong correlation between `neomor` and `un5mor` (0.928) emphasizes that neonatal deaths contribute significantly to overall under-five mortality rates.

Mean total deaths in Western Asia over the years



The trend in the plot reflects a fluctuating but overall sharp increase in mean total deaths in Western Asia from 2000 to 2019, with significant spikes in 2013 through 2016. Deaths remained relatively low from 2000 to 2004, but a notable increase occurred from 2005 onwards, peaking in 2015. This trend likely correlates with major conflicts in the region, including the U.S. invasion of Iraq (2003), the Syrian Civil War (starting in 2011), the rise of ISIS (particularly between 2014-2016), and the Yemeni Civil War (intensifying in 2015). These conflicts, along with heightened sectarian violence, external interventions, and the breakdown of state structures, could have driven the high death tolls during these years.

Conclusion and Next Steps

In this exploratory data analysis, we identified several key patterns and relationships within the dataset. Initial visualizations and summary statistics provided insights into the distributions of the variables, the presence of outliers, and potential correlations among features. The findings suggest significant variability in some key indicators, such as mortality rates, and highlight areas that may warrant further investigation. However, we still need to perform several important steps. Moving forward, we will focus on: - Feature Selection: Applying techniques to identify the most relevant predictors for the target variable(s), reducing dimensionality, and improving model performance. - Statistical Tests: Conducting formal hypothesis tests to assess the statistical significance of observed patterns and relationships. - Modeling: Exploring appropriate statistical models, like like Random Forest Classifier, XGBoost, or Regression models, to predict outcomes or further explain the relationships in the data.