ArmedConflictEDA

Xinze Yu

Study Objectives and Operationalizations

The **objective** of the current analysis is to study how armed conflicts impact the maternal and child health from a global scope.

Specifically, the primary **exposure variable** of interest is **armed conflict**, as defined by the UCDP, it is a a binary variable indicating the presence of conflict for each country–year observation (0 = no, $<$ 25 battle-related deaths; 1 = yes, $\leq$ 25 battle-related deaths).

The primary **outcome** measures are maternal, under-5, infant, and neonatal **mortality rates**.

A list of covariates is included in the dataset, and will potentially be included in the model: "gdp1000", "OECD", "OECD2023", "popdens", "urban", "agedep", "male_edu", "temp", "rainfall1000", "Drought", "Earthquake".

1. Explore Data Structure and Summary Statistics of Key Variables

We can start from checking the overall structure, data type, and missing values.

```
# import data
acdata <- read.csv(here('data', 'analytical', 'finaldata.csv'), header = TRUE)
# factorize primary binary exposure
acdata$armed_conflict <- as.factor(acdata$armed_conflict)

# check structure, summary statistics, and missingness
str(acdata)
```

```
'data.frame':   3720 obs. of  21 variables:
 $ country_name  : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ ISO           : chr  "AFG" "AFG" "AFG" "AFG" ...
 $ region        : chr  "Southern Asia" "Southern Asia" "Southern Asia" "Southern Asia" ...
 $ Year          : int  2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 ...
 $ gdp1000       : num  NA NA 0.184 0.2 0.222 ...
 $ OECD          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ OECD2023      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ popdens       : num  14.1 14.2 14.3 14.4 15.2 ...
 $ urban         : num  16.3 16.3 16.4 16.6 16.7 ...
 $ agedep        : num  108 109 109 109 109 ...
 $ male_edu      : num  2.76 2.86 2.95 3.05 3.16 ...
 $ temp          : num  12.7 12.9 12.7 12.2 13 ...
 $ rainfall1000  : num  0.276 0.279 0.381 0.429 0.375 ...
 $ MatMortality  : int  1450 1390 1300 1240 1180 1140 1120 1090 1030 993 ...
 $ InfMortality  : num  90.5 87.9 85.3 82.7 80 77.3 74.6 71.9 69.2 66.7 ...
 $ NeoMortality  : num  60.9 59.7 58.5 57.2 55.9 54.6 53.2 51.7 50.3 48.9 ...
 $ Und5Mortality : num  129 125 121 117 113 ...
 $ totdeath      : int  5065 5394 5553 1157 944 817 1711 4982 7020 5660 ...
 $ armed_conflict: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ Drought       : int  1 0 0 0 0 0 1 0 1 0 ...
 $ Earthquake    : int  0 1 1 1 1 1 1 0 0 1 ...
```

```
summary(acdata)
```

```
 country_name           ISO               region               Year
 Length:3720        Length:3720        Length:3720        Min.   :2000
 Class :character   Class :character   Class :character   1st Qu.:2005
 Mode  :character   Mode  :character   Mode  :character   Median :2010
                                                          Mean   :2010
                                                          3rd Qu.:2014
                                                          Max.   :2019

    gdp1000              OECD          OECD2023         popdens
 Min.   :  0.1105   Min.   :0.000   Min.   :0.0000   Min.   : 0.00
 1st Qu.:  1.2383   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:14.79
 Median :  4.0719   Median :0.000   Median :0.0000   Median :27.52
 Mean   : 11.4917   Mean   :0.171   Mean   :0.1882   Mean   :30.57
 3rd Qu.: 13.1531   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:40.72
 Max.   :123.6787   Max.   :1.000   Max.   :1.0000   Max.   :99.86
 NA's   :62                                          NA's   :20
     urban             agedep          male_edu           temp
```

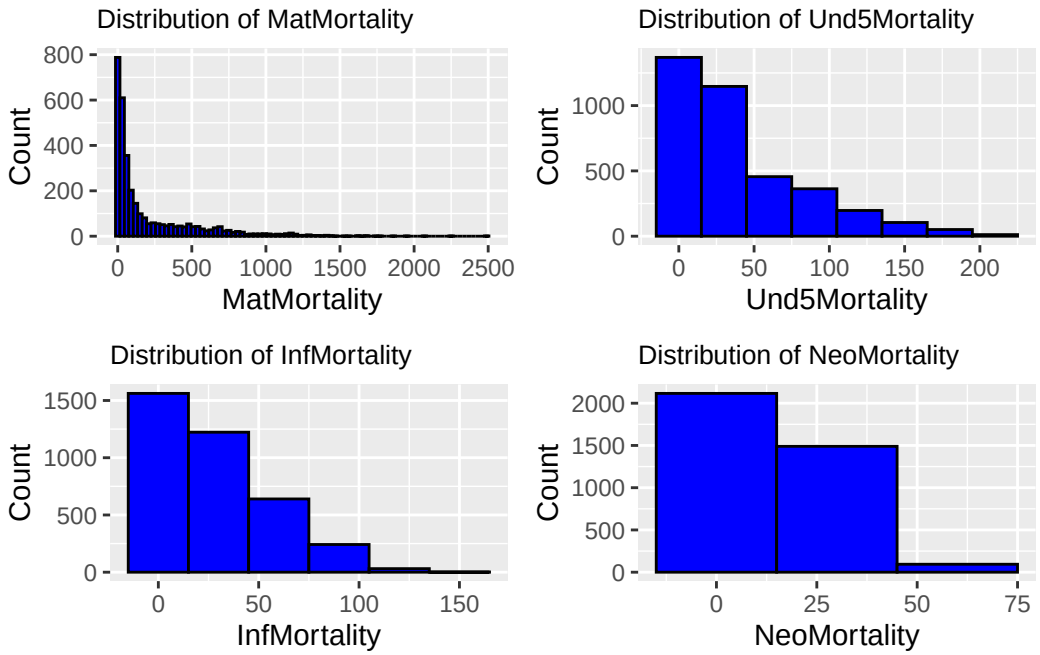
```
Min.   : 0.1025   Min.   : 16.17   Min.   : 1.067   Min.   :-2.405
1st Qu.:17.2872   1st Qu.: 47.94   1st Qu.: 5.904   1st Qu.:12.928
Median :30.2535   Median : 55.51   Median : 8.368   Median :21.958
Mean   :30.6948   Mean   : 61.94   Mean   : 8.258   Mean   :19.625
3rd Qu.:41.6558   3rd Qu.: 77.11   3rd Qu.:10.849   3rd Qu.:25.869
Max.   :93.4135   Max.   :111.48   Max.   :14.441   Max.   :29.676
NA's   :20                         NA's   :20       NA's   :20
 rainfall1000      MatMortality     InfMortality     NeoMortality
Min.   :0.01993   Min.   :   2.0   Min.   :  1.60   Min.   : 0.80
1st Qu.:0.59146   1st Qu.:  17.0   1st Qu.:  7.60   1st Qu.: 4.90
Median :1.01288   Median :  66.0   Median : 18.90   Median :12.10
Mean   :1.20216   Mean   : 210.6   Mean   : 28.90   Mean   :16.18
3rd Qu.:1.68706   3rd Qu.: 299.8   3rd Qu.: 44.52   3rd Qu.:25.32
Max.   :4.71081   Max.   :2480.0   Max.   :138.10   Max.   :60.90
NA's   :20        NA's   :426      NA's   :20       NA's   :20
Und5Mortality       totdeath       armed_conflict    Drought
Min.   :  2.00   Min.   :    0.0   0:3016         Min.   :0.00000
1st Qu.:  9.00   1st Qu.:    0.0   1: 704         1st Qu.:0.00000
Median : 22.20   Median :    0.0                  Median :0.00000
Mean   : 40.50   Mean   :  361.1                  Mean   :0.08737
3rd Qu.: 61.33   3rd Qu.:    2.0                  3rd Qu.:0.00000
Max.   :224.90   Max.   :78644.0                  Max.   :1.00000
NA's   :20
  Earthquake
Min.   :0.00000
1st Qu.:0.00000
Median :0.00000
Mean   :0.08333
3rd Qu.:0.00000
Max.   :1.00000
```

Note that one of the key outcome variables, maternal mortality, has noticeable missingness - 426 (11.5%) observations were missing out of 3720 observations.

2. Visualize the Distributions

Plot **histograms** for outcome variables to identify skewness or outliers.

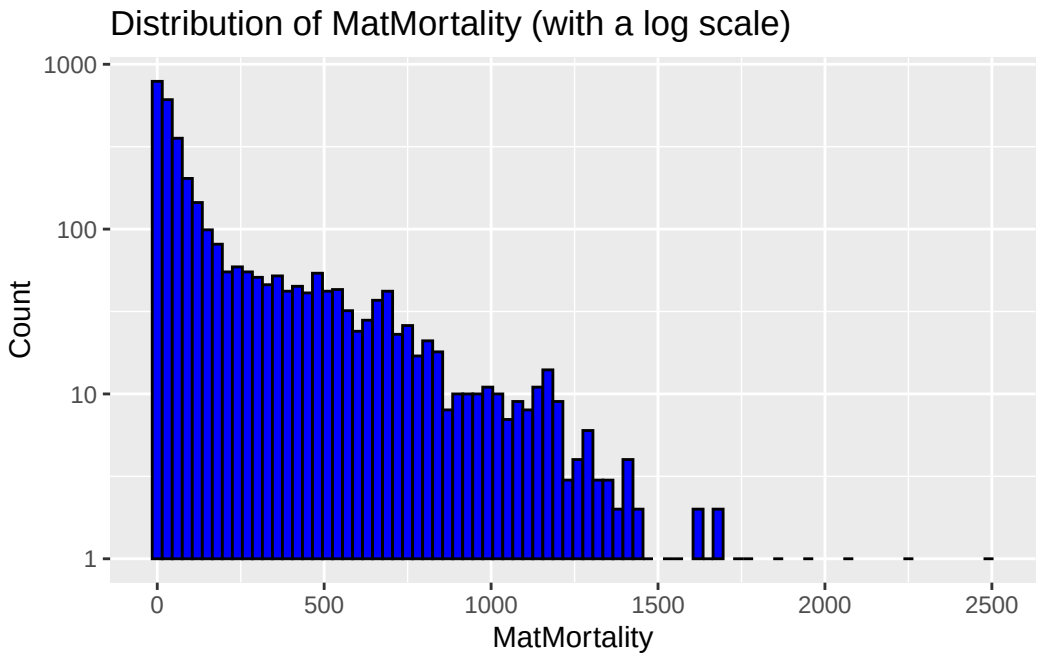
```r
# define outcome
outcomes <- c("MatMortality", "Und5Mortality", "InfMortality", "NeoMortality")

# create empty list to store plots
plot_list <- list()

# create histogram for each outcome
for (var in outcomes) {
  p <- ggplot(acdata, aes_string(x = var)) +
    geom_histogram(binwidth = 30, fill = "blue", color = "black") +
    labs(title = paste("Distribution of", var), x = var, y = "Count") +
    theme(plot.title = element_text(size = 10))
  plot_list[[var]] <- p
}

# arrange plots into grid layout
grid.arrange(grobs = plot_list, ncol = 2)
```



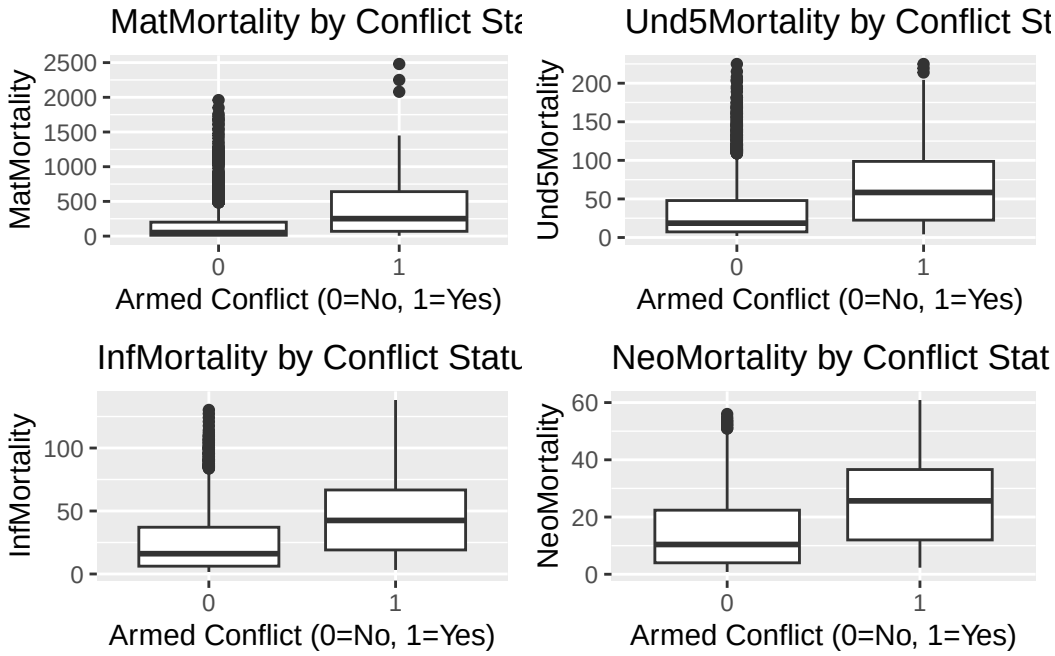
The largely empty graph in maternal mortality suggests the presence of outliers. We can proceed to use logarithms.

```r
# create histogram for maternal mortality with log scale
ggplot(acdata, aes_string(x = "MatMortality")) +
```

```
    geom_histogram(binwidth = 30, fill = "blue", color = "black") +
    scale_y_log10() +
    labs(title = paste("Distribution of", "MatMortality", "(with a log scale)"),
         x = "MatMortality", y = "Count")
```




```
# create empty list to store plots
plot_list1 <- list()

# create boxplot to compare each outcome between
# conflict and non-conflict countries
for (var in outcomes) {
  p <- ggplot(acdata, aes_string(x = "armed_conflict", y = var)) +
    geom_boxplot() +
    labs(title = paste(var, "by Conflict Status"),
         x = "Armed Conflict (0=No, 1=Yes)", y = var)
  plot_list1[[var]] <- p
}

# arrange plots into grid layout
grid.arrange(grobs = plot_list1, ncol = 2)
```

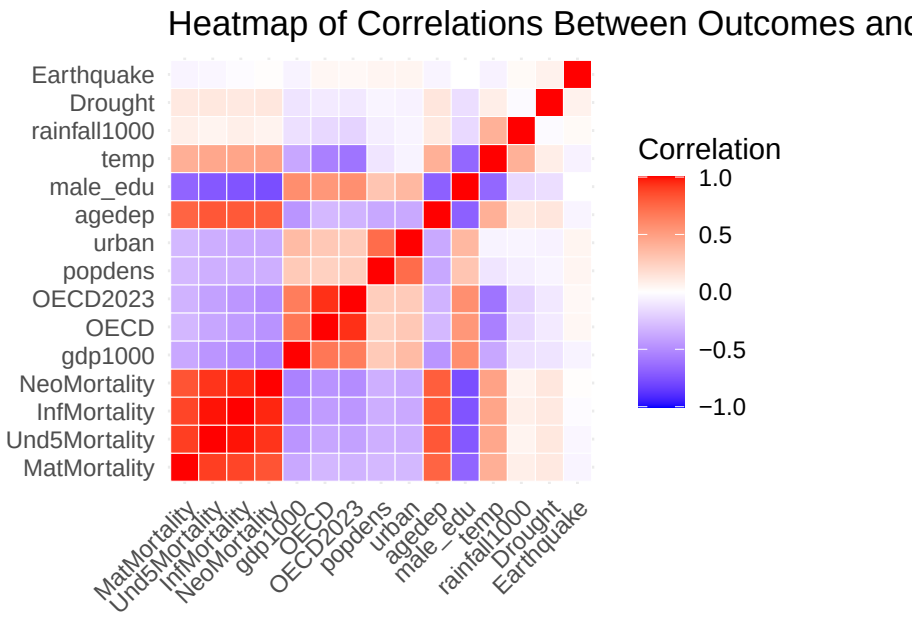


Overall, the country/year with armed conflicts shows higher mortality rates among all outcome groups.

3. Explore Relationships between Variables

```
# define covariates
covariates <- c("gdp1000", "OECD", "OECD2023", "popdens",
                "urban", "agedep", "male_edu", "temp",
                "rainfall1000", "Drought", "Earthquake")

# correlation between key variable
cor.mat <- cor(acdata[, c(outcomes, covariates)], use = "complete.obs")

# melt the correlation matrix to long format for ggplot
melted.cor.mat <- melt(cor.mat)

# Draw the heatmap using ggplot
ggplot(data = melted.cor.mat, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
```

```
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  coord_fixed() +
  labs(title = "Heatmap of Correlations Between Outcomes and Covariates",
       x = "", y = "")
```



Summary of the EDA on Armed Conflict and Maternal & Child Health

1. Data Structure and Summary

The dataset contains 3,720 observations across 21 variables, representing multiple countries over the period 2000–2019. Key variables include maternal mortality, infant mortality, and under-5 mortality, which show insignificant levels of missingness, with maternal mortality having highest rate of missingness (11.5%) that might be concerning.

2. Distribution of Key Variables

Histograms of the outcome variables reveal highly skewed distributions with potential outliers for maternal mortality. To manage this skewness, a log scale transformation was applied to the maternal mortality data, which provided a clearer view of the distribution.

Boxplots comparing conflict vs. non-conflict countries indicate higher mortality rates for maternal, under-5, infant, and neonatal mortality in conflict-affected countries. The presence of conflict appears to show a noticable association with worsened health outcomes for mothers and children.

3. Correlations and Relationships

A correlation heatmap was created to examine the relationships between the outcome variables and covariates. Key findings include: strong negative correlations between male_edu and all mortality measures; strong positive correlations between all mortality rates and agedep and temp. No significant correlations between natural disasters (e.g., drought, earthquakes) and the mortality outcomes.