CHL8010: Statistical Programming and Computation in Health Data

Belina

2024-09-30

Exploratory data analysis

```
# load the data
final_data <- read.csv(here("data/analytical", "final_data.csv"), header = TRUE)
```

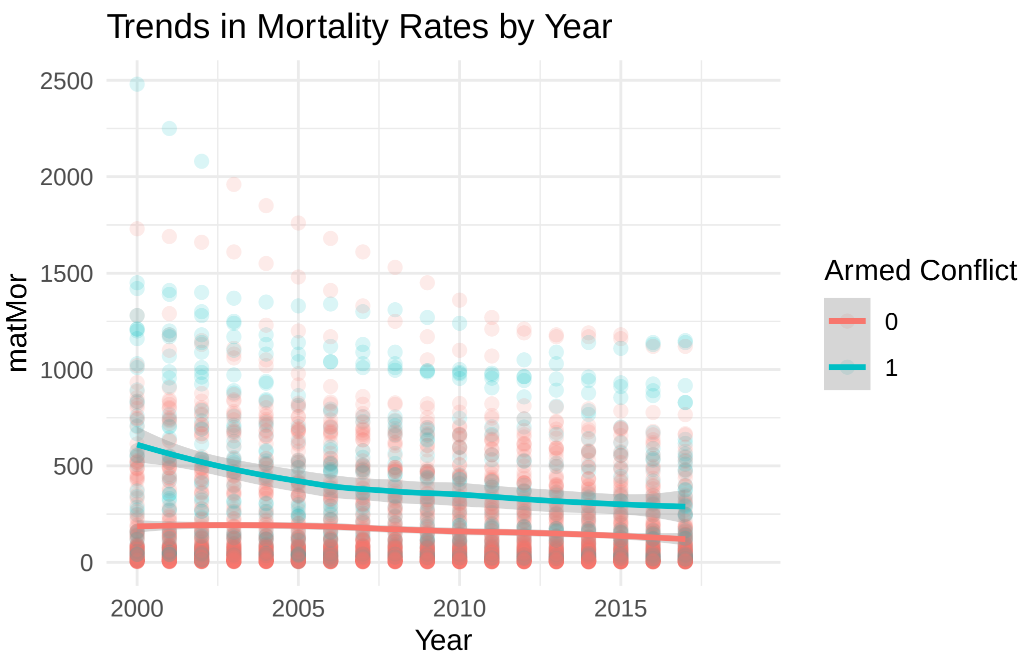
```
glimpse(final_data)
```

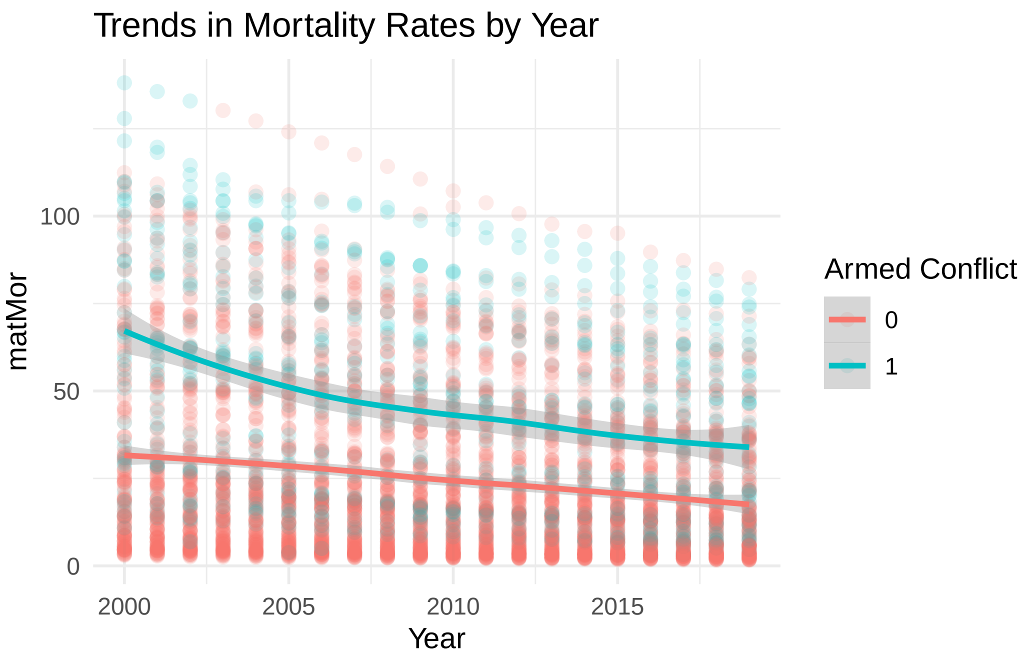
```
Rows: 3,720
Columns: 21
$ country_name   <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanist~
$ ISO            <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG",~
$ region         <chr> "Southern Asia", "Southern Asia", "Southern Asia", "Sou~
$ Year           <int> 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2~
$ gdp1000        <dbl> NA, NA, 0.1835328, 0.2004626, 0.2216576, 0.2550551, 0.2~
$ OECD           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ OECD2023       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ popdens        <dbl> 14.13654, 14.23156, 14.32270, 14.40691, 15.21947, 15.33~
$ urban          <dbl> 16.25324, 16.25661, 16.42654, 16.60701, 16.71367, 16.85~
$ agedep         <dbl> 108.34663, 108.98989, 109.34716, 109.44753, 109.28682, ~
$ male_edu       <dbl> 2.762086, 2.856936, 2.954241, 3.054121, 3.156706, 3.262~
$ temp           <dbl> 12.69959, 12.85570, 12.71081, 12.16592, 13.04643, 12.23~
$ rainfall1000   <dbl> 0.2763704, 0.2793079, 0.3805710, 0.4288939, 0.3754336, ~
$ totaldeath     <int> 5065, 5394, 5553, 1157, 944, 817, 1711, 4982, 7020, 566~
$ armed_conflict <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ Earthquake     <int> 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0~
$ Drought        <int> 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1~
$ matMor         <int> 1450, 1390, 1300, 1240, 1180, 1140, 1120, 1090, 1030, 9~
$ infMor         <dbl> 90.5, 87.9, 85.3, 82.7, 80.0, 77.3, 74.6, 71.9, 69.2, 6~
$ neoMor         <dbl> 60.9, 59.7, 58.5, 57.2, 55.9, 54.6, 53.2, 51.7, 50.3, 4~
$ under5Mor      <dbl> 129.2, 125.2, 121.1, 116.9, 112.6, 108.4, 104.1, 99.9, ~
```

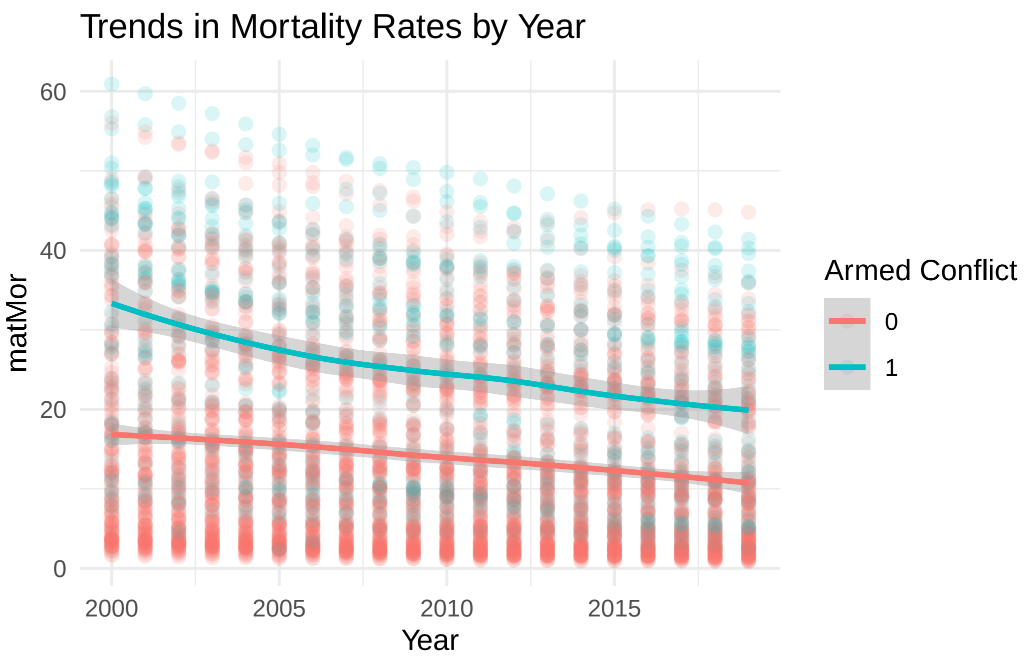
```r
summary(final_data)
```

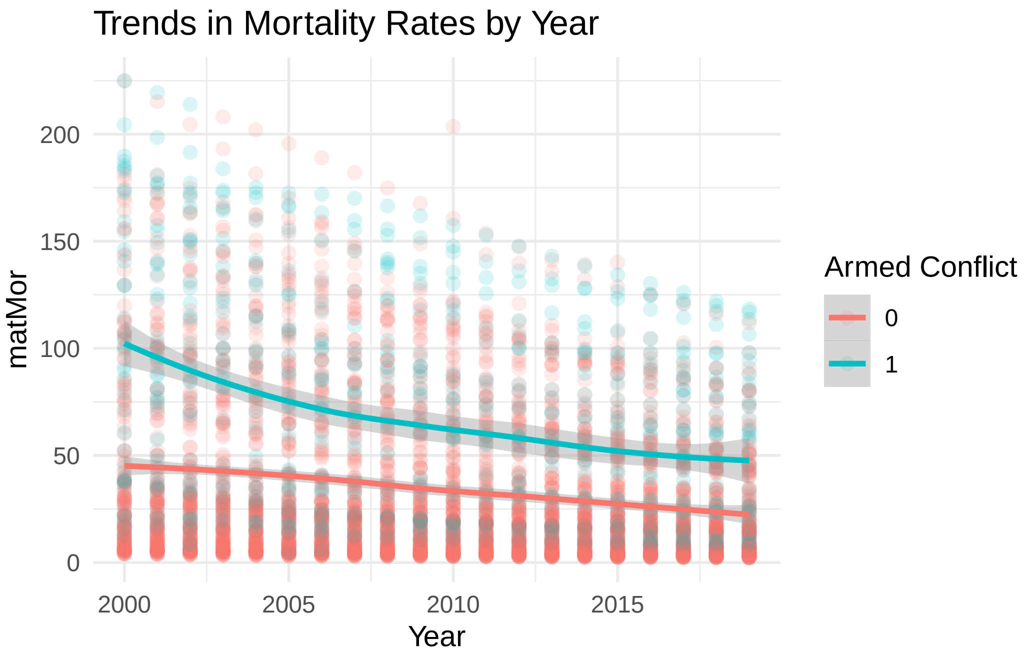
```
 country_name           ISO               region                Year     
 Length:3720        Length:3720        Length:3720        Min.   :2000  
 Class :character   Class :character   Class :character   1st Qu.:2005  
 Mode  :character   Mode  :character   Mode  :character   Median :2010  
                                                          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       totaldeath      armed_conflict     Earthquake     
 Min.   :0.01993   Min.   :    0.0   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:0.59146   1st Qu.:    0.0   1st Qu.:0.0000   1st Qu.:0.00000  
 Median :1.01288   Median :    0.0   Median :0.0000   Median :0.00000  
 Mean   :1.20216   Mean   :  361.1   Mean   :0.1892   Mean   :0.08333  
 3rd Qu.:1.68706   3rd Qu.:    2.0   3rd Qu.:0.0000   3rd Qu.:0.00000  
 Max.   :4.71081   Max.   :78644.0   Max.   :1.0000   Max.   :1.00000  
 NA's   :20                                                            
    Drought             matMor           infMor           neoMor     
 Min.   :0.00000   Min.   :   2.0   Min.   :  1.60   Min.   : 0.80  
 1st Qu.:0.00000   1st Qu.:  17.0   1st Qu.:  7.60   1st Qu.: 4.90  
 Median :0.00000   Median :  66.0   Median : 18.90   Median :12.10  
 Mean   :0.08737   Mean   : 210.6   Mean   : 28.90   Mean   :16.18  
 3rd Qu.:0.00000   3rd Qu.: 299.8   3rd Qu.: 44.52   3rd Qu.:25.32  
 Max.   :1.00000   Max.   :2480.0   Max.   :138.10   Max.   :60.90  
                   NA's   :426      NA's   :20       NA's   :20     
   under5Mor     
 Min.   :  2.00  
 1st Qu.:  9.00  
 Median : 22.20  
 Mean   : 40.50  
 3rd Qu.: 61.33  
 Max.   :224.90  
 NA's   :20      
```

```r
Mor_names <- c("matMor","infMor","neoMor","under5Mor")

for (Mor in Mor_names){
  plot <- final_data %>% ggplot(aes(Year,.data[[Mor]],
                                    color=factor(armed_conflict))) +
    geom_point(alpha=0.15, size=2) + geom_smooth(aes(group = armed_conflict),
                                                 method = "loess") +
    labs(
      title = "Trends in Mortality Rates by Year",
      x = "Year",
      y = Mor_names,
      color = "Armed Conflict"
    ) +
    theme_minimal()
  print(plot)
}
```

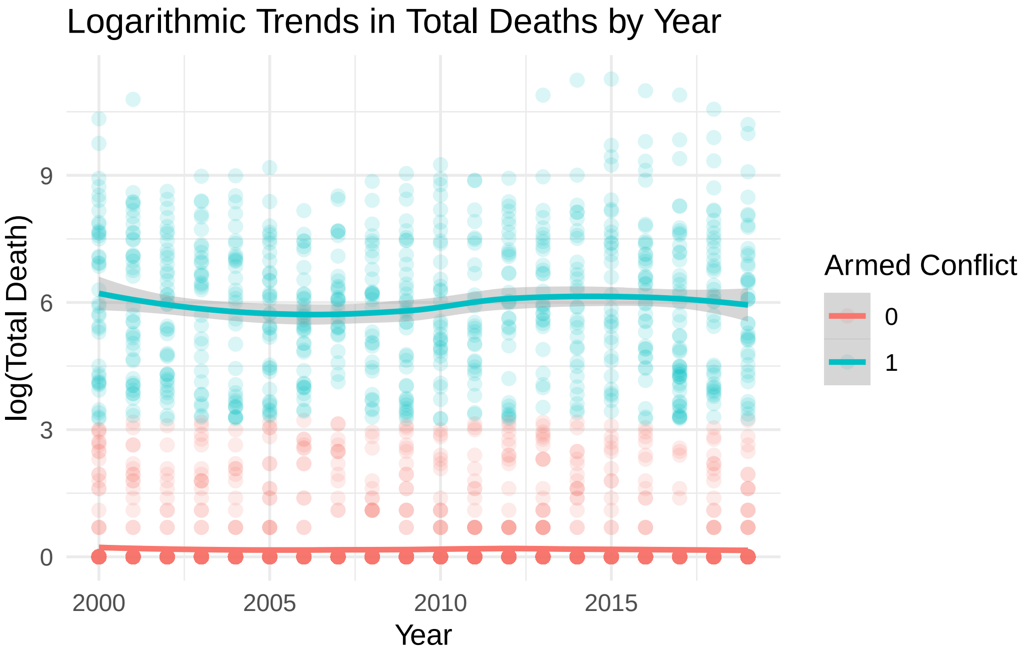

Trends in Mortality Rates by Year
matMor
2500
2000
1500
1000
500
0
2000
2005
2010
2015
Year
Armed Conflict
0
1


Trends in Mortality Rates by Year
matMor
100
50
0
2000
2005
2010
2015
Year
Armed Conflict
0
1





```
final_data %>% ggplot(aes(Year,log(totaldeath+1), color=factor(armed_conflict))) +
    geom_point(alpha=0.15, size=2) + geom_smooth(aes(group = armed_conflict),
                                                  method = "loess") +
    labs(
      title = "Logarithmic Trends in Total Deaths by Year",
      x = "Year",
      y = "log(Total Death)",
      color = "Armed Conflict"
    ) + theme_minimal()
```



Investigate Final Data without countries with 3 highest total death counts for each year.

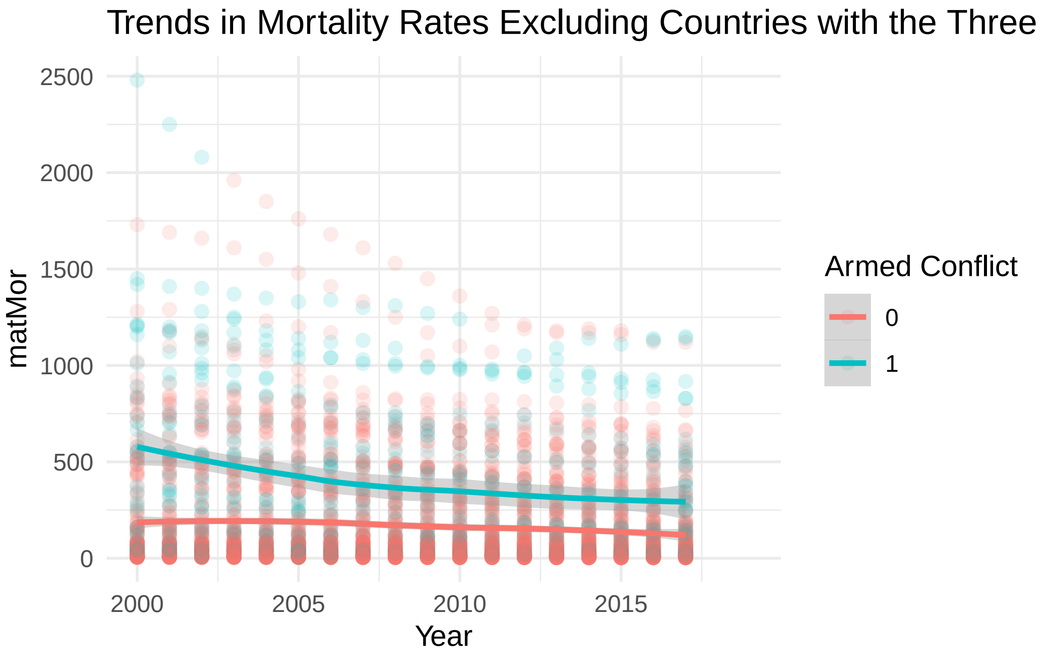
```
highest_totaldeath <- final_data %>%
  group_by(Year) %>%
  slice_max(totaldeath, n = 3)

highest_totaldeath %>% select(country_name,Year, totaldeath)
```

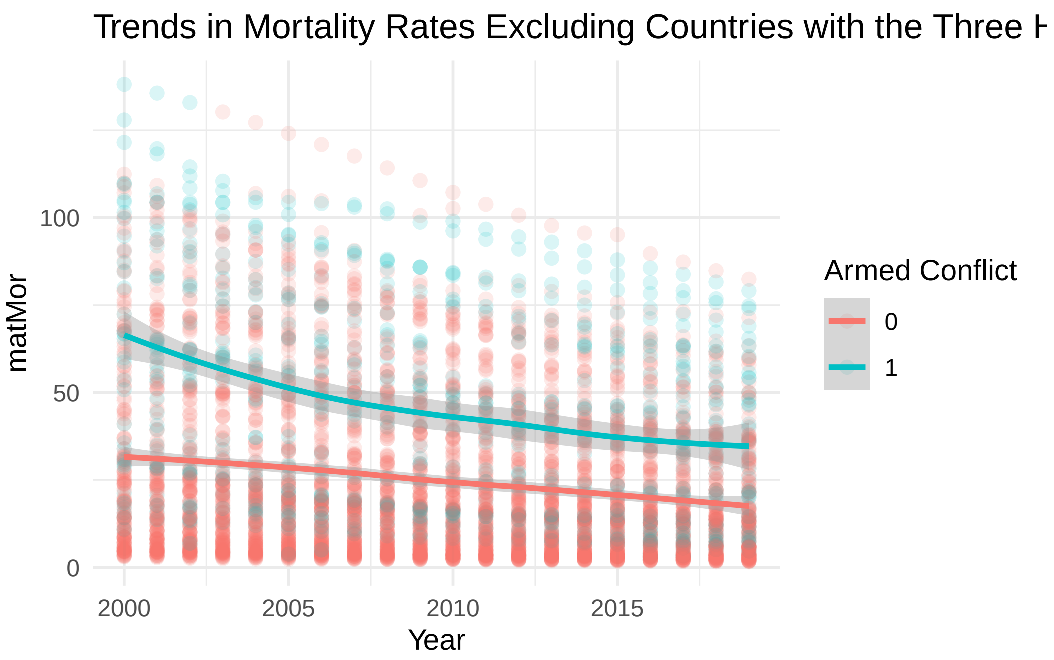
```
# A tibble: 60 x 3
# Groups:   Year [20]
   country_name                       Year totaldeath
   <chr>                             <int>      <int>
 1 Ethiopia                           2000      30786
 2 Eritrea                            2000      17203
 3 Democratic Republic of the Congo  2000       7541
 4 Ethiopia                          2001      48666
 5 Afghanistan                       2001       5394
 6 Russian Federation                2001       4333
 7 Afghanistan                       2002       5553
 8 Colombia                          2002       4592
 9 Sudan                             2002       3719
10 Democratic Republic of the Congo  2003       7931
# i 50 more rows
```

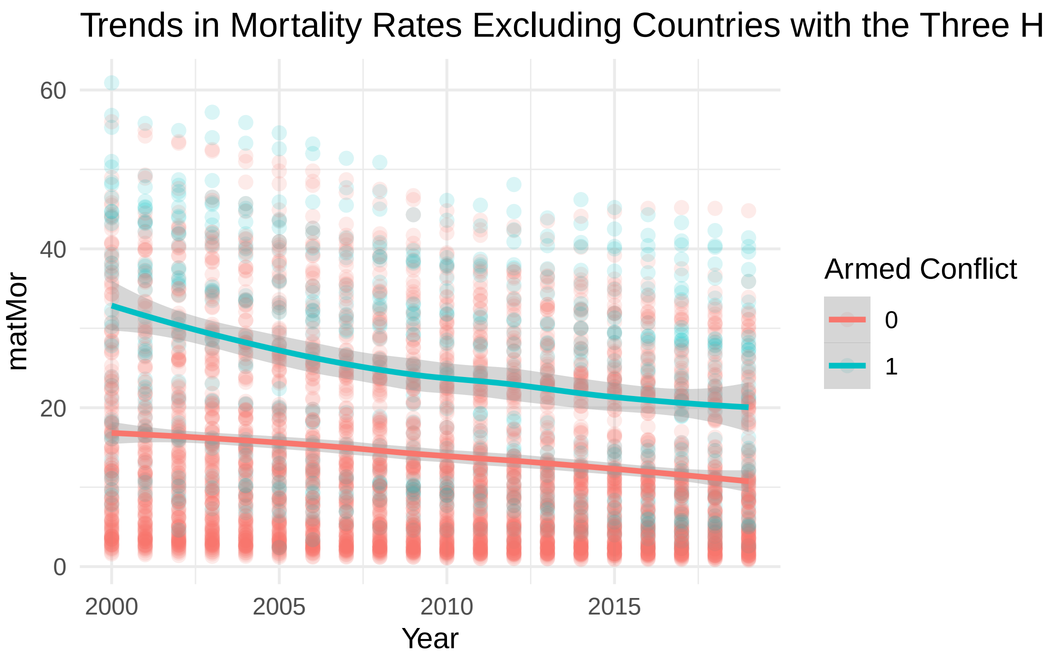
```
finaldata_without3highest <- anti_join(final_data,highest_totaldeath,
                                       by=c("country_name","Year"))

head(finaldata_without3highest)
```

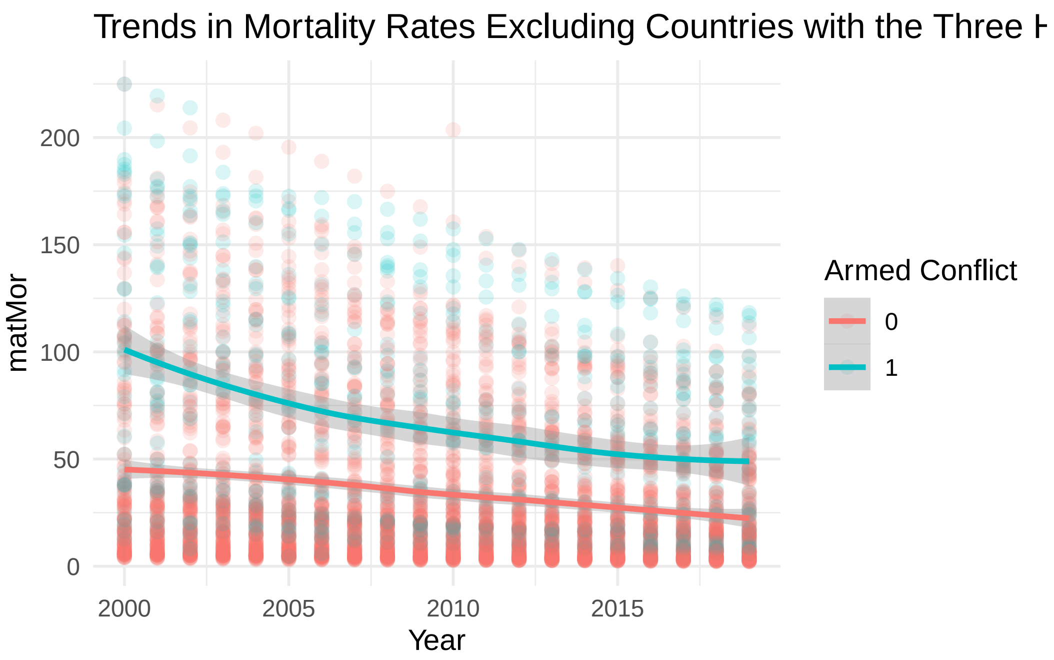
```
  country_name ISO          region Year   gdp1000 OECD OECD2023  popdens
1  Afghanistan AFG   Southern Asia 2000        NA    0        0 14.13654
2  Afghanistan AFG   Southern Asia 2003 0.2004626    0        0 14.40691
3  Afghanistan AFG   Southern Asia 2004 0.2216576    0        0 15.21947
4  Afghanistan AFG   Southern Asia 2005 0.2550551    0        0 15.33619
5  Afghanistan AFG   Southern Asia 2006 0.2740005    0        0 15.43982
6      Albania ALB Southern Europe 2000 1.1266833    0        0 33.08368
     urban   agedep male_edu     temp rainfall1000 totaldeath armed_conflict
1 16.25324 108.3466 2.762086 12.69959    0.2763704       5065              1
2 16.60701 109.4475 3.054121 12.16592    0.4288939       1157              1
3 16.71367 109.2868 3.156706 13.04643    0.3754336        944              1
4 16.85096 107.9646 3.262133 12.23141    0.4415680        817              1
5 16.98105 106.3262 3.370551 12.96153    0.4437097       1711              1
6 27.38836  59.6573 8.961755 13.73920    0.7971749          6              0
  Earthquake Drought matMor infMor neoMor under5Mor
1          0       1   1450   90.5   60.9     129.2
2          1       0   1240   82.7   57.2     116.9
3          1       0   1180   80.0   55.9     112.6
4          1       0   1140   77.3   54.6     108.4
5          1       1   1120   74.6   53.2     104.1
6          0       0     23   24.1   12.1      27.2
```

```
Mor_names <- c("matMor","infMor","neoMor","under5Mor")

for (Mor in Mor_names){
  plot <- finaldata_without3highest %>% ggplot(aes(Year,.data[[Mor]],
                                                   color=factor(armed_conflict))) +
    geom_point(alpha=0.15, size=2) + geom_smooth(aes(group = armed_conflict),
                                                 method = "loess") +
    labs(
      title = "Trends in Mortality Rates Excluding Countries with the Three Highest Total Dea
      x = "Year",
      y = Mor_names,
      color = "Armed Conflict"
    ) +
    theme_minimal()
  print(plot)
}
```

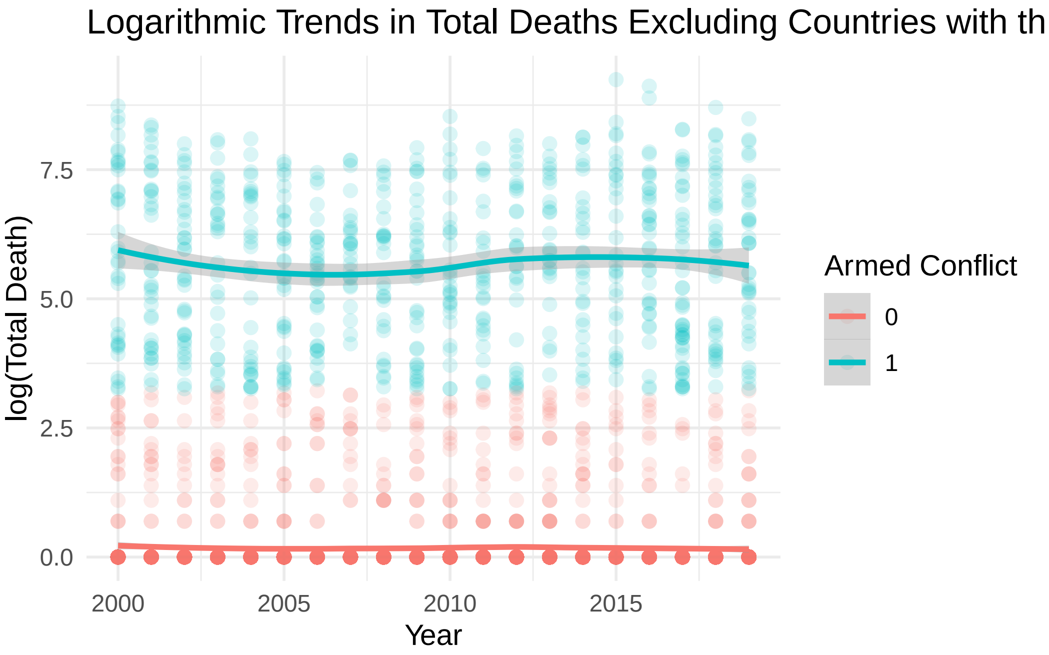



Trends in Mortality Rates Excluding Countries with the Three H


Trends in Mortality Rates Excluding Countries with the Three H

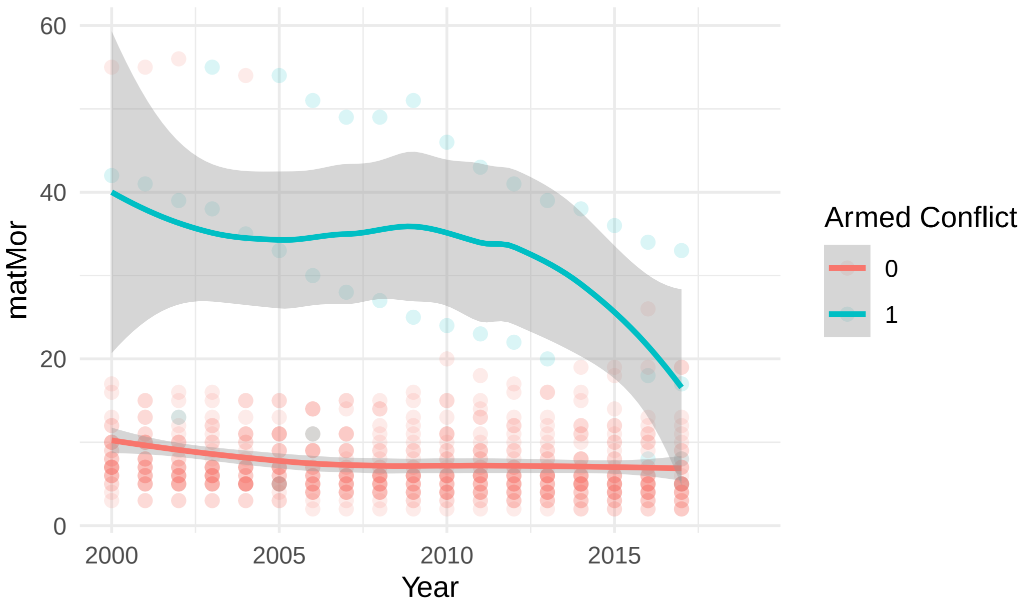


```
finaldata_without3highest %>% ggplot(aes(Year,log(totaldeath+1),
                                        color=factor(armed_conflict))) +
    geom_point(alpha=0.15, size=2) + geom_smooth(aes(group = armed_conflict),
                                                method = "loess") +
    labs(
      title = "Logarithmic Trends in Total Deaths Excluding Countries with the Three Highest
      x = "Year",
      y = "log(Total Death)",
      color = "Armed Conflict"
    ) + theme_minimal()
```

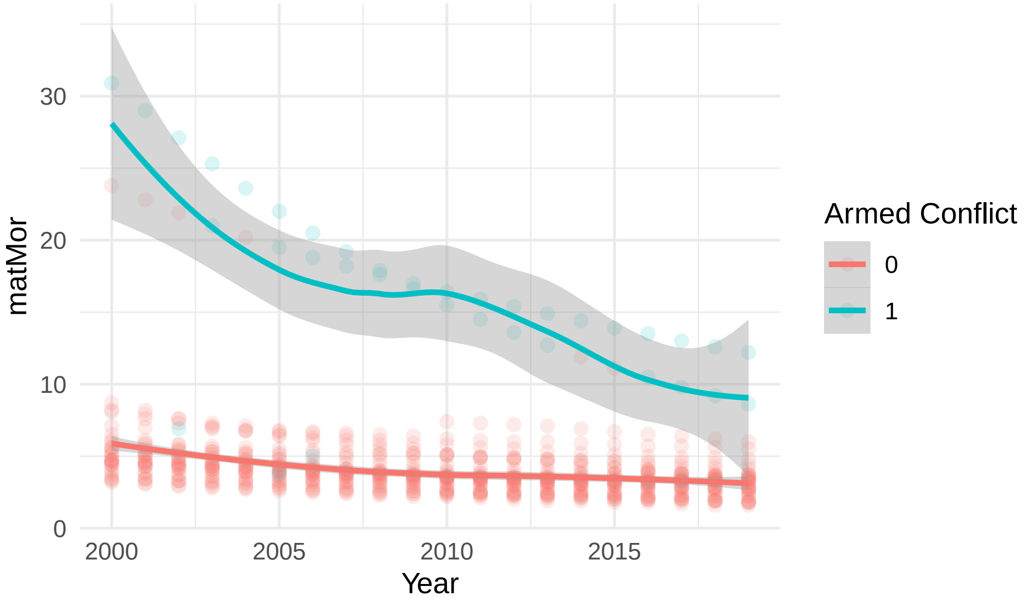


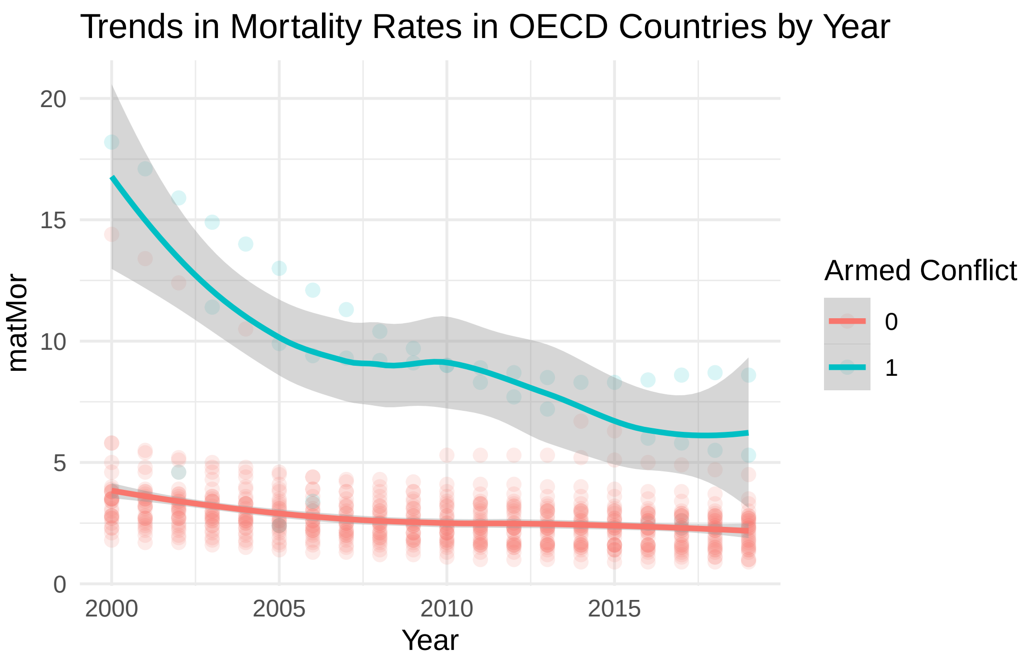
```
oecd_countries <- final_data %>%
  filter(OECD==1)

oecd_conflict <- oecd_countries %>% filter(armed_conflict==1)

Mor_names <- c("matMor","infMor","neoMor","under5Mor")

for (Mor in Mor_names){
  plot <- oecd_countries %>% ggplot(aes(Year,.data[[Mor]],
                                        color=factor(armed_conflict))) +
    geom_point(alpha=0.15, size=2) + geom_smooth(aes(group = armed_conflict),
                                                 method = "loess") +
    labs(
      title = "Trends in Mortality Rates in OECD Countries by Year",
      x = "Year",
      y = Mor_names,
      color = "Armed Conflict"
    ) +
    theme_minimal()
  print(plot)
}
```

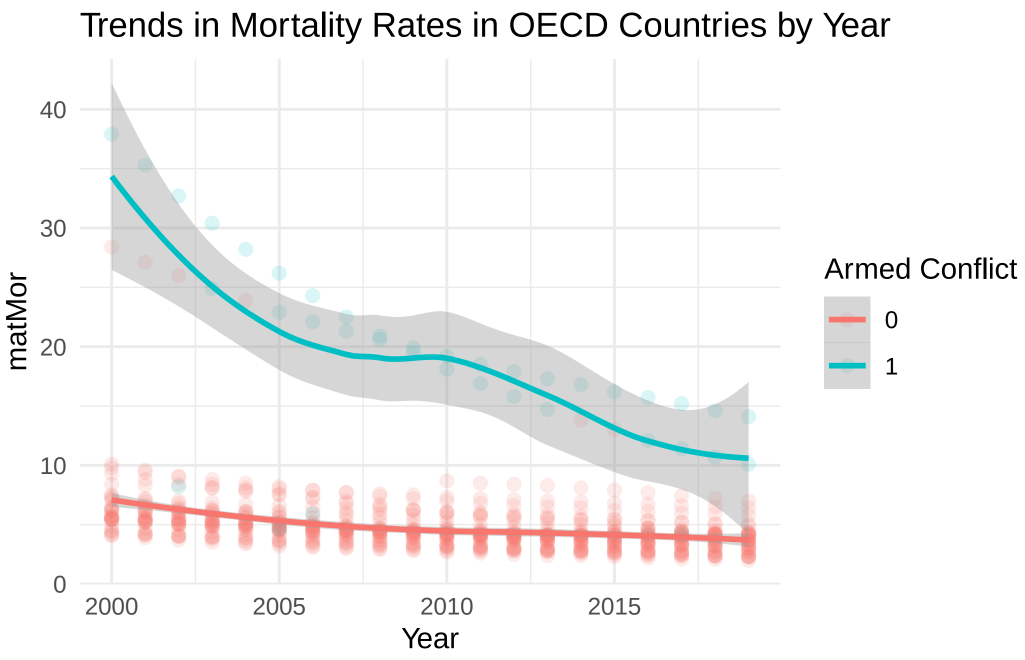
Trends in Mortality Rates in OECD Countries by Year



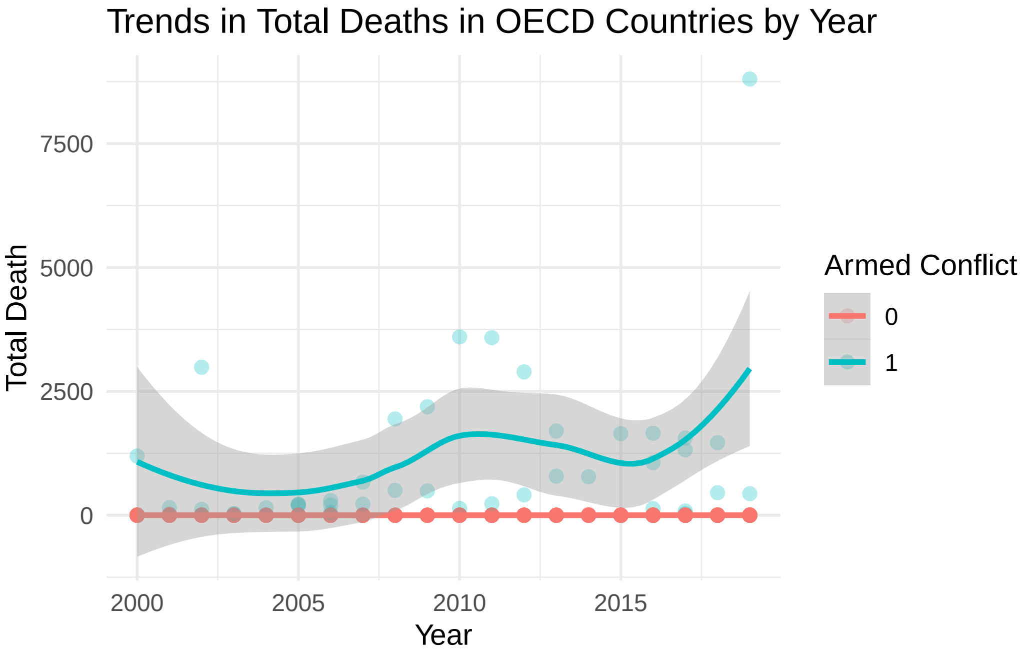
Trends in Mortality Rates in OECD Countries by Year







```
oecd_countries %>% ggplot(aes(Year,totaldeath, color=factor(armed_conflict))) +
    geom_point(alpha=0.3, size=2) + geom_smooth(aes(group = armed_conflict),
                                               method = "loess") +
    labs(
    title = "Trends in Total Deaths in OECD Countries by Year",
    x = "Year",
    y = "Total Death",
    color = "Armed Conflict"
  ) + theme_minimal()
```



Summary