Exploratory Data Analysis

Meagan Lacroix

```
library(here)
```

```
Warning: package 'here' was built under R version 4.3.3
```

```
library(dplyr)
library(arrow)
```

```
Warning: package 'arrow' was built under R version 4.3.3
```

```
library(janitor)
```

```
Warning: package 'janitor' was built under R version 4.3.3
```

```
library(knitr)
```

```
Warning: package 'knitr' was built under R version 4.3.2
```

```
library(lubridate)
```

```
Warning: package 'lubridate' was built under R version 4.3.2
```

```
library(mice)
```

```
Warning: package 'mice' was built under R version 4.3.3
```

```
library(modelsummary)
```

```
Warning: package 'modelsummary' was built under R version 4.3.3
```

```
library(naniar)
```

```
Warning: package 'naniar' was built under R version 4.3.3
```

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.2
```

Read in the merged data

```
eda_data <- read.csv(here("data", "finaldata.csv"), header = TRUE)
```

Check names, head/tail

```
names(eda_data)
```

```
 [1] "country_name" "ISO"          "region"       "year"         "gdp1000"
 [6] "OECD"         "OECD2023"     "popdens"      "urban"        "agedep"
[11] "male_edu"     "temp"         "rainfall1000" "totdeath"     "armconf1"
[16] "drought"      "earthquake"   "Mor_inf"      "Mor_neonat"   "Mor_under5"
[21] "Mor_mat"
```

```
head(eda_data)
```

```
  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 drought earthquake
1 108.3466 2.762086 12.69959    0.2763704     5065        1       1          0
2 108.9899 2.856936 12.85570    0.2793079     5394        1       0          1
3 109.3472 2.954241 12.71081    0.3805710     5553        1       0          1
4 109.4475 3.054121 12.16592    0.4288939     1157        1       0          1
5 109.2868 3.156706 13.04643    0.3754336      944        1       0          1
6 107.9646 3.262133 12.23141    0.4415680      817        1       0          1
  Mor_inf Mor_neonat Mor_under5 Mor_mat
1    90.5       60.9      129.2    1450
2    87.9       59.7      125.2    1390
3    85.3       58.5      121.1    1300
4    82.7       57.2      116.9    1240
5    80.0       55.9      112.6    1180
6    77.3       54.6      108.4    1140
```

```
tail(eda_data)
```

```
     country_name ISO             region year  gdp1000 OECD OECD2023  popdens
3715     Zimbabwe ZWE Sub-Saharan Africa 2014 1.407034    0        0 26.52884
3716     Zimbabwe ZWE Sub-Saharan Africa 2015 1.410329    0        0 26.54454
3717     Zimbabwe ZWE Sub-Saharan Africa 2016 1.421788    0        0 26.53811
3718     Zimbabwe ZWE Sub-Saharan Africa 2017 1.192107    0        0 26.49281
3719     Zimbabwe ZWE Sub-Saharan Africa 2018 2.269177    0        0 26.47943
3720     Zimbabwe ZWE Sub-Saharan Africa 2019 1.421869    0        0 26.46341
        urban   agedep male_edu     temp rainfall1000 totdeath armconf1 drought
3715 24.40427 85.87550 8.679591 20.87651    0.6777257        0        0       0
3716 24.75233 85.08337 8.785078 21.45470    0.4490721        0        0       0
3717 25.02842 84.11222 8.889947 21.39290    0.4939246        0        0       0
3718 25.29333 83.10129 8.994252 20.85962    0.9533149        0        0       1
3719 25.53759 82.12335 9.098048 20.86041    0.9535655        0        0       0
3720 25.70572 81.20786 9.201384 20.86120    0.9538138        4        0       0
     earthquake Mor_inf Mor_neonat Mor_under5 Mor_mat
3715          0    42.9       28.2       62.7     494
3716          0    42.1       27.8       61.3     480
3717          0    40.8       27.4       58.7     468
3718          0    39.9       27.0       57.0     458
3719          0    38.8       26.6       54.8      NA
3720          0    38.1       26.2       54.2      NA
```

Creating more consistent variable names

```
eda_data <- eda_data %>%
  clean_names()
```

Changing OECD, OECD2023, armconf1, drought, and earthquate to a factor

```
eda_data <- eda_data %>%
  mutate(across(c(oecd, oecd2023, armconf1, drought, earthquake), as.factor))
```

Summarize the variables. There is a large amount of missing data for maternal mortality. This will be investigated further

```
eda_data %>% summary()
```

```
 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           urban
 Min.   :  0.1105   0:3084   0:3020   Min.   : 0.00   Min.   : 0.1025
 1st Qu.:  1.2383   1: 636   1: 700   1st Qu.:14.79   1st Qu.:17.2872
 Median :  4.0719                     Median :27.52   Median :30.2535
 Mean   : 11.4917                     Mean   :30.57   Mean   :30.6948
 3rd Qu.: 13.1531                     3rd Qu.:40.72   3rd Qu.:41.6558
 Max.   :123.6787                     Max.   :99.86   Max.   :93.4135
 NA's   :62                           NA's   :20      NA's   :20
     agedep          male_edu           temp          rainfall1000
 Min.   : 16.17   Min.   : 1.067   Min.   :-2.405   Min.   :0.01993
 1st Qu.: 47.94   1st Qu.: 5.904   1st Qu.:12.928   1st Qu.:0.59146
 Median : 55.51   Median : 8.368   Median :21.958   Median :1.01288
 Mean   : 61.94   Mean   : 8.258   Mean   :19.625   Mean   :1.20216
 3rd Qu.: 77.11   3rd Qu.:10.849   3rd Qu.:25.869   3rd Qu.:1.68706
```

```
Max.   :111.48   Max.   :14.441   Max.   :29.676   Max.   :4.71081
                 NA's   :20       NA's   :20       NA's   :20
   totdeath       armconf1 drought  earthquake    mor_inf
Min.   :    0.0   0:3016   0:3395   0:3410     Min.   :  1.60
1st Qu.:    0.0   1: 704   1: 325   1: 310     1st Qu.:  7.60
Median :    0.0                                Median : 18.90
Mean   :  361.1                                Mean   : 28.90
3rd Qu.:    2.0                                3rd Qu.: 44.52
Max.   :78644.0                                Max.   :138.10
                                               NA's   :20
  mor_neonat      mor_under5        mor_mat
Min.   : 0.80   Min.   :  2.00   Min.   :   2.0
1st Qu.: 4.90   1st Qu.:  9.00   1st Qu.:  17.0
Median :12.10   Median : 22.20   Median :  66.0
Mean   :16.18   Mean   : 40.50   Mean   : 210.6
3rd Qu.:25.32   3rd Qu.: 61.33   3rd Qu.: 299.8
Max.   :60.90   Max.   :224.90   Max.   :2480.0
NA's   :20      NA's   :20       NA's   :426
```

Checking the names of the character variables. There are 186 countries and 17 regions

```
unique(eda_data$country_name)
```

```
  [1] "Afghanistan"                    "Albania"
  [3] "Algeria"                        "Andorra"
  [5] "Angola"                         "Antigua and Barbuda"
  [7] "Argentina"                      "Armenia"
  [9] "Australia"                      "Austria"
 [11] "Azerbaijan"                     "Bahrain"
 [13] "Bangladesh"                     "Barbados"
 [15] "Belarus"                        "Belgium"
 [17] "Belize"                         "Benin"
 [19] "Bhutan"                         "Bolivia"
 [21] "Bosnia and Herzegovina"         "Botswana"
 [23] "Brazil"                         "Brunei"
 [25] "Bulgaria"                       "Burkina Faso"
 [27] "Burundi"                        "Cambodia"
 [29] "Cameroon"                       "Canada"
 [31] "Cape Verde"                     "Central African Republic"
 [33] "Chad"                           "Chile"
```

```
 [35] "China"                              "Colombia"
 [37] "Comoros"                            "Congo"
 [39] "Costa Rica"                         "Cote d'Ivoire"
 [41] "Croatia"                            "Cuba"
 [43] "Cyprus"                             "Czech Republic"
 [45] "Democratic Republic of the Congo" "Denmark"
 [47] "Djibouti"                           "Dominica"
 [49] "Dominican Republic"                 "Ecuador"
 [51] "Egypt"                              "El Salvador"
 [53] "Equatorial Guinea"                  "Eritrea"
 [55] "Estonia"                            "Ethiopia"
 [57] "Federated States of Micronesia"   "Fiji"
 [59] "Finland"                            "France"
 [61] "Gabon"                              "Georgia"
 [63] "Germany"                            "Ghana"
 [65] "Greece"                             "Grenada"
 [67] "Guatemala"                          "Guinea"
 [69] "Guinea-Bissau"                      "Guyana"
 [71] "Haiti"                              "Honduras"
 [73] "Hungary"                            "Iceland"
 [75] "India"                              "Indonesia"
 [77] "Iran"                               "Iraq"
 [79] "Ireland"                            "Italy"
 [81] "Jamaica"                            "Japan"
 [83] "Jordan"                             "Kazakhstan"
 [85] "Kenya"                              "Kiribati"
 [87] "Kuwait"                             "Kyrgyzstan"
 [89] "Laos"                               "Latvia"
 [91] "Lebanon"                            "Lesotho"
 [93] "Liberia"                            "Libya"
 [95] "Lithuania"                          "Luxembourg"
 [97] "Macedonia"                          "Madagascar"
 [99] "Malawi"                             "Malaysia"
[101] "Maldives"                           "Mali"
[103] "Malta"                              "Marshall Islands"
[105] "Mauritania"                         "Mauritius"
[107] "Mexico"                             "Moldova"
[109] "Mongolia"                           "Montenegro"
[111] "Morocco"                            "Mozambique"
[113] "Myanmar"                            "Namibia"
[115] "Nepal"                              "Netherlands"
[117] "New Zealand"                        "Nicaragua"
[119] "Niger"                              "Nigeria"
```

```
[121] "North Korea"                    "Norway"
[123] "Oman"                           "Pakistan"
[125] "Panama"                         "Papua New Guinea"
[127] "Paraguay"                       "Peru"
[129] "Philippines"                    "Poland"
[131] "Portugal"                       "Puerto Rico"
[133] "Qatar"                          "Romania"
[135] "Russian Federation"             "Rwanda"
[137] "Saint Lucia"                    "Saint Vincent and the Grenadines"
[139] "Samoa"                          "Sao Tome and Principe"
[141] "Saudi Arabia"                   "Senegal"
[143] "Serbia"                         "Seychelles"
[145] "Sierra Leone"                   "Singapore"
[147] "Slovakia"                       "Slovenia"
[149] "Solomon Islands"                "Somalia"
[151] "South Africa"                   "South Korea"
[153] "South Sudan"                    "Spain"
[155] "Sri Lanka"                      "Sudan"
[157] "Suriname"                       "Swaziland"
[159] "Sweden"                         "Switzerland"
[161] "Syria"                          "Tajikistan"
[163] "Tanzania"                       "Thailand"
[165] "The Bahamas"                    "The Gambia"
[167] "Timor-Leste"                    "Togo"
[169] "Tonga"                          "Trinidad and Tobago"
[171] "Tunisia"                        "Turkey"
[173] "Turkmenistan"                   "Uganda"
[175] "Ukraine"                        "United Arab Emirates"
[177] "United Kingdom"                 "United States"
[179] "Uruguay"                        "Uzbekistan"
[181] "Vanuatu"                        "Venezuela"
[183] "Vietnam"                        "Yemen"
[185] "Zambia"                         "Zimbabwe"
```

```
unique(eda_data$iso)
```

```
  [1] "AFG" "ALB" "DZA" "AND" "AGO" "ATG" "ARG" "ARM" "AUS" "AUT" "AZE" "BHR"
 [13] "BGD" "BRB" "BLR" "BEL" "BLZ" "BEN" "BTN" "BOL" "BIH" "BWA" "BRA" "BRN"
 [25] "BGR" "BFA" "BDI" "KHM" "CMR" "CAN" "CPV" "CAF" "TCD" "CHL" "CHN" "COL"
 [37] "COM" "COG" "CRI" "CIV" "HRV" "CUB" "CYP" "CZE" "COD" "DNK" "DJI" "DMA"
 [49] "DOM" "ECU" "EGY" "SLV" "GNQ" "ERI" "EST" "ETH" "FSM" "FJI" "FIN" "FRA"
 [61] "GAB" "GEO" "DEU" "GHA" "GRC" "GRD" "GTM" "GIN" "GNB" "GUY" "HTI" "HND"
```

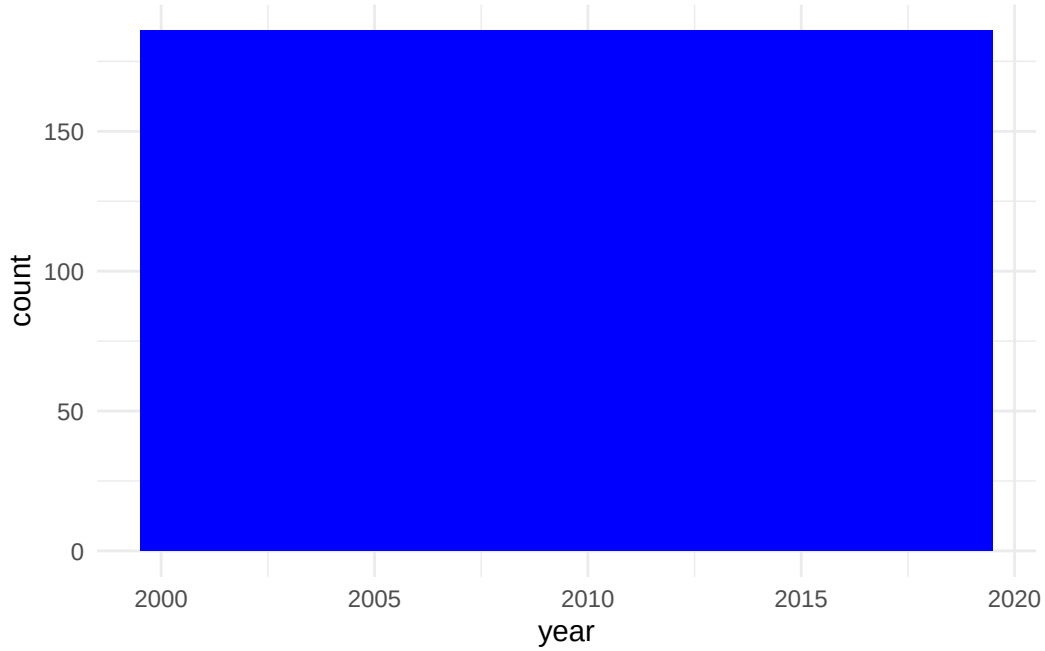
```
 [73] "HUN" "ISL" "IND" "IDN" "IRN" "IRQ" "IRL" "ITA" "JAM" "JPN" "JOR" "KAZ"
 [85] "KEN" "KIR" "KWT" "KGZ" "LAO" "LVA" "LBN" "LSO" "LBR" "LBY" "LTU" "LUX"
 [97] "MKD" "MDG" "MWI" "MYS" "MDV" "MLI" "MLT" "MHL" "MRT" "MUS" "MEX" "MDA"
[109] "MNG" "MNE" "MAR" "MOZ" "MMR" "NAM" "NPL" "NLD" "NZL" "NIC" "NER" "NGA"
[121] "PRK" "NOR" "OMN" "PAK" "PAN" "PNG" "PRY" "PER" "PHL" "POL" "PRT" "PRI"
[133] "QAT" "ROU" "RUS" "RWA" "LCA" "VCT" "WSM" "STP" "SAU" "SEN" "SRB" "SYC"
[145] "SLE" "SGP" "SVK" "SVN" "SLB" "SOM" "ZAF" "KOR" "SSD" "ESP" "LKA" "SDN"
[157] "SUR" "SWZ" "SWE" "CHE" "SYR" "TJK" "TZA" "THA" "BHS" "GMB" "TLS" "TGO"
[169] "TON" "TTO" "TUN" "TUR" "TKM" "UGA" "UKR" "ARE" "GBR" "USA" "URY" "UZB"
[181] "VUT" "VEN" "VNM" "YEM" "ZMB" "ZWE"
```

```r
unique(eda_data$region)
```

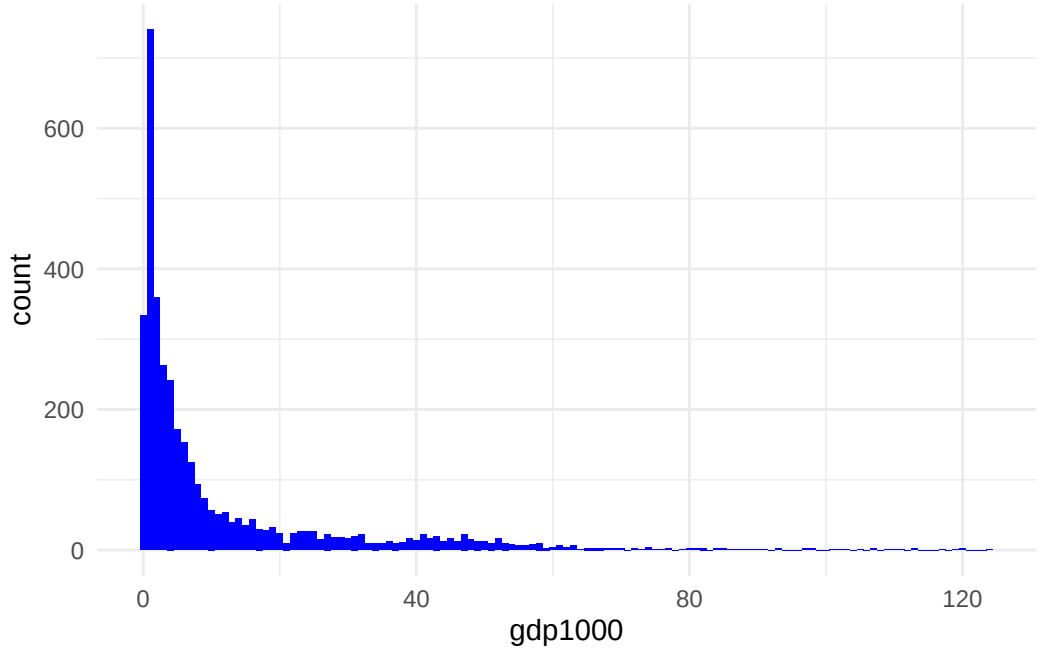
```
 [1] "Southern Asia"                   "Southern Europe"
 [3] "Northern Africa"                 "Sub-Saharan Africa"
 [5] "Latin America and the Caribbean" "Western Asia"
 [7] "Australia and New Zealand"       "Western Europe"
 [9] "Eastern Europe"                  "South-eastern Asia"
[11] "Northern America"                "Eastern Asia"
[13] "Northern Europe"                 "Micronesia"
[15] "Melanesia"                       "Central Asia"
[17] "Polynesia"
```

Creating histograms of the numeric variables. Total deaths is difficult to visualize because there is a large number of 0-1 deaths and a very small frequency of larger # deaths. This will be investigated further.

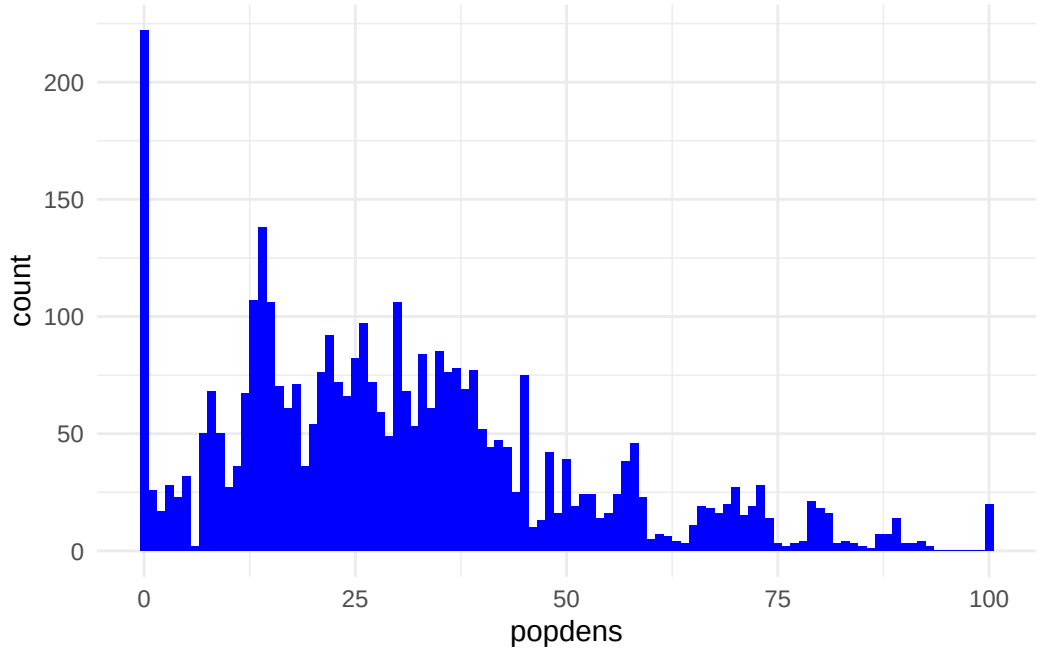
```r
numeric_vars <- eda_data %>%
  select(where(is.numeric)) %>%
  names()

for (var in numeric_vars) {
  p <- ggplot(eda_data, aes(x = !!sym(var))) +  # Dynamically assign variable
    geom_histogram(binwidth = 1, fill = "blue") +
    xlab(var) +  # Set x-axis label to the variable name
    theme_minimal()

  print(p)
}
```



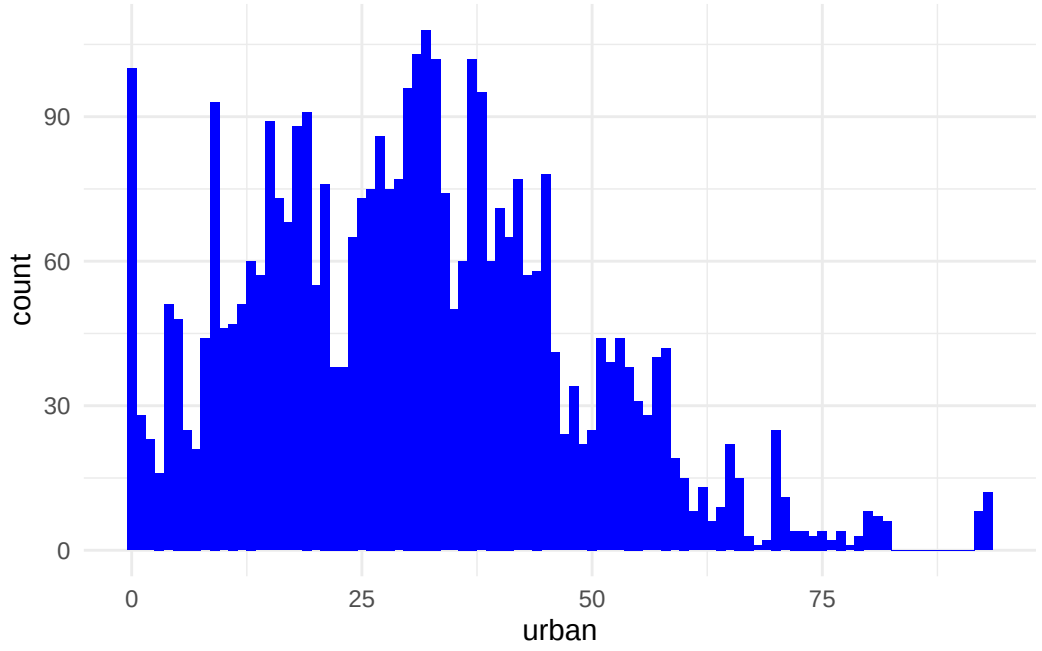

```
Warning: Removed 62 rows containing non-finite values (`stat_bin()`).
```

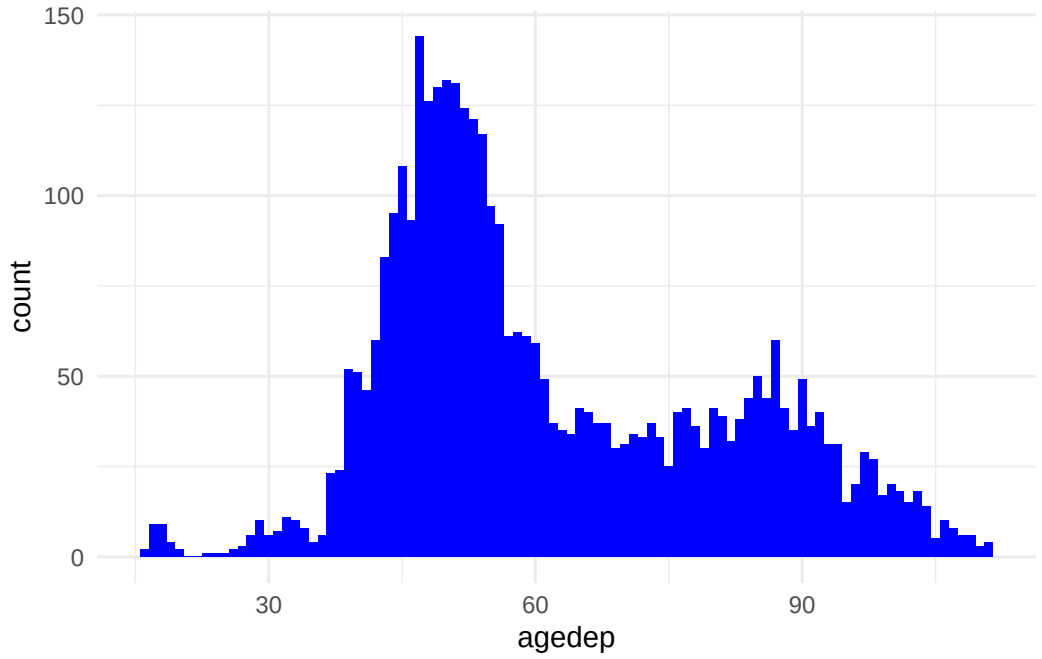



```
Warning: Removed 20 rows containing non-finite values (`stat_bin()`).
```

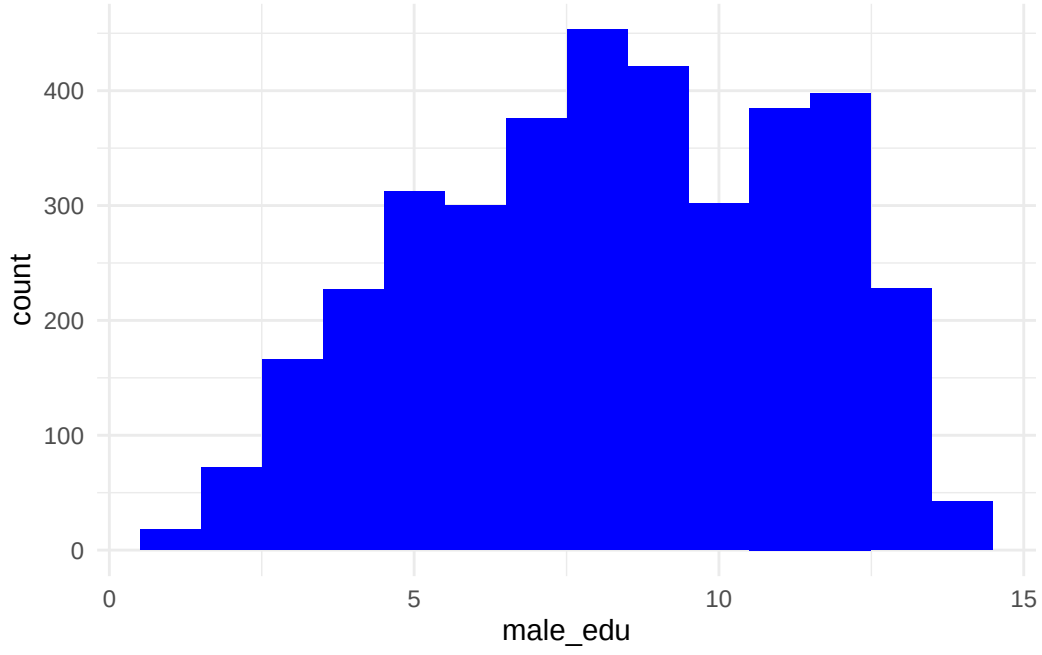



```
Warning: Removed 20 rows containing non-finite values (`stat_bin()`).
```

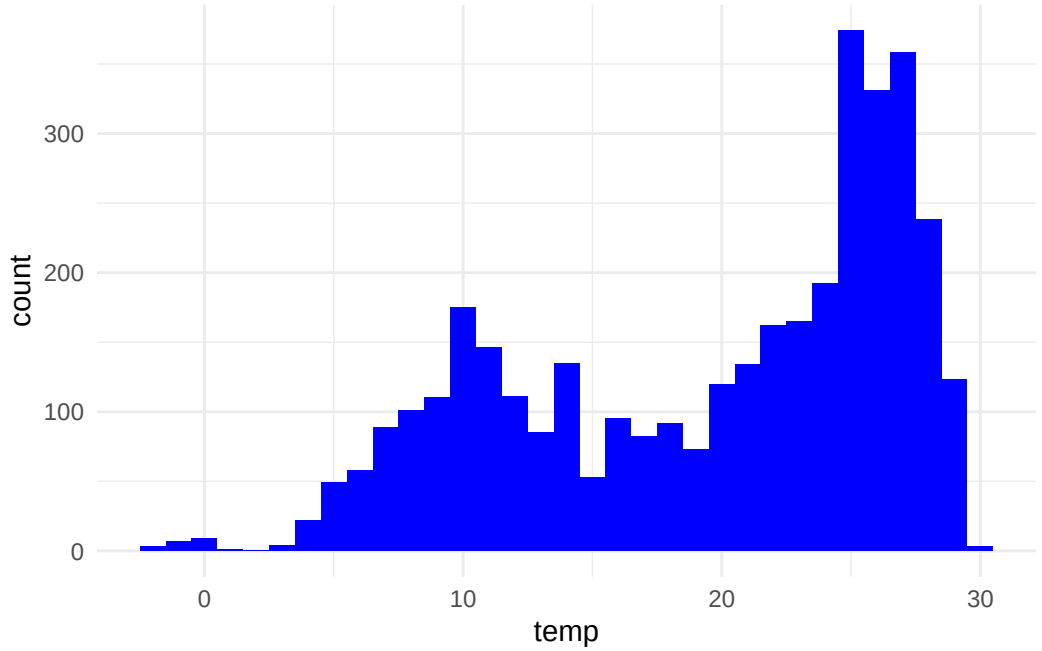




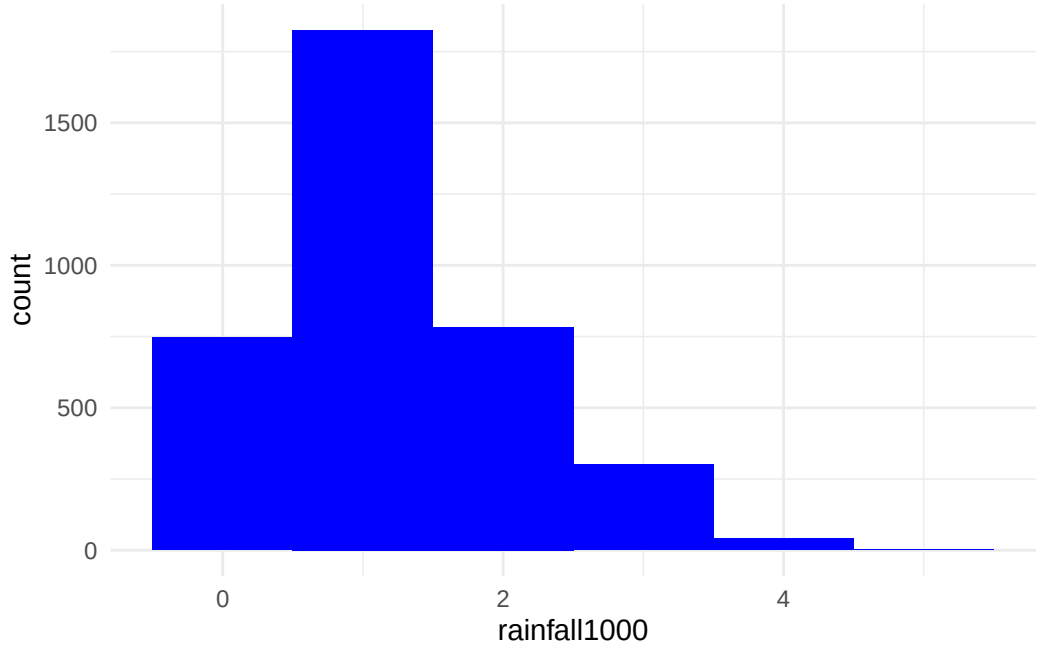

```
Warning: Removed 20 rows containing non-finite values (`stat_bin()`).
```

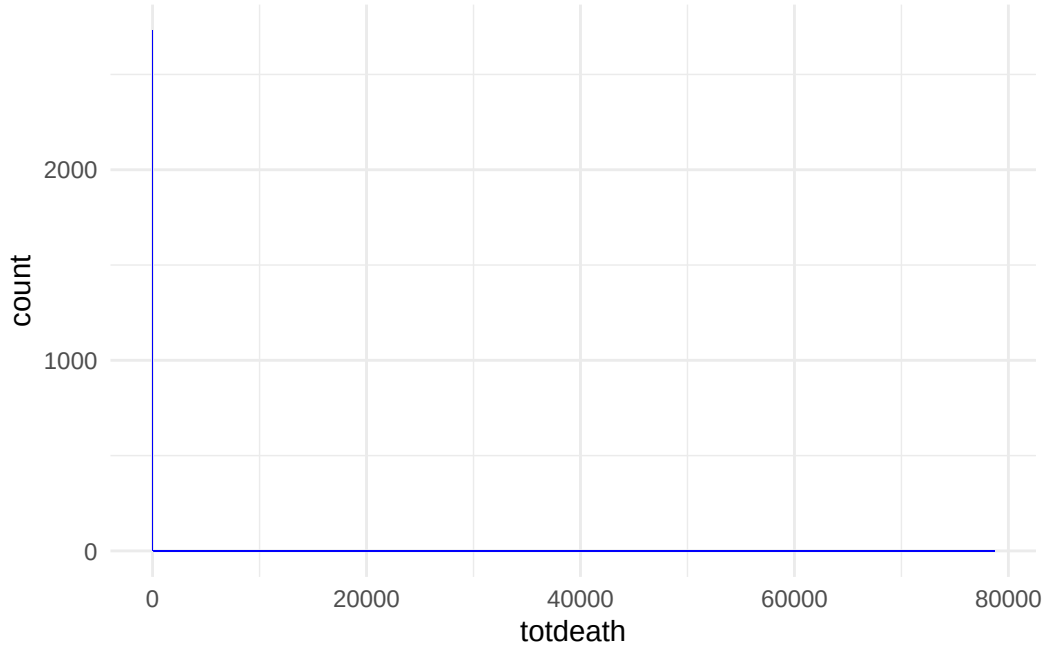



```
Warning: Removed 20 rows containing non-finite values (`stat_bin()`).
```

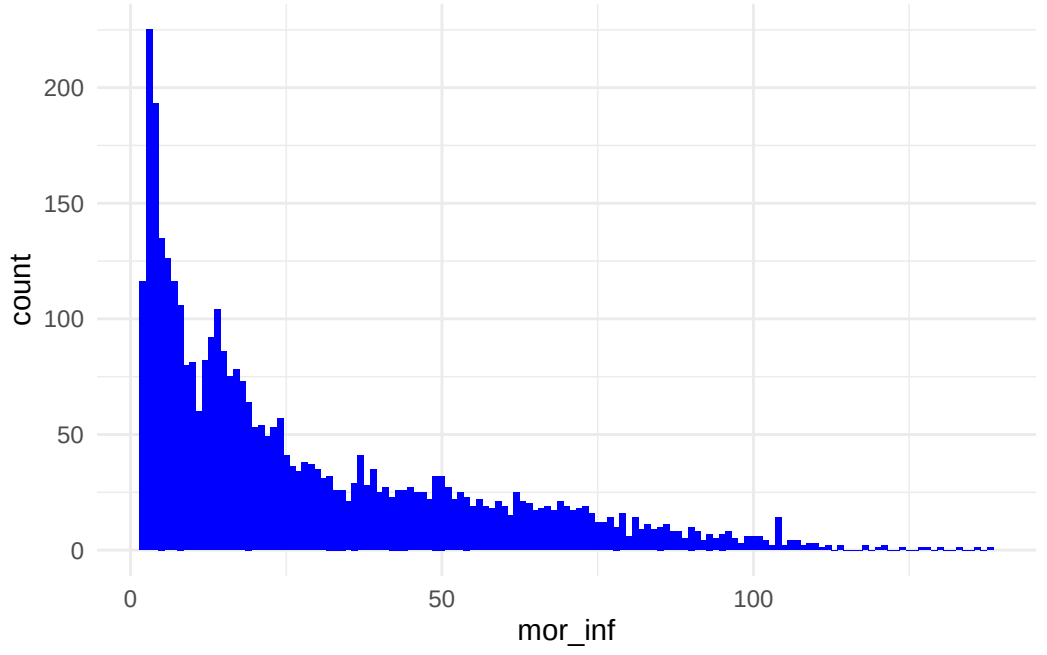



```
Warning: Removed 20 rows containing non-finite values (`stat_bin()`).
```

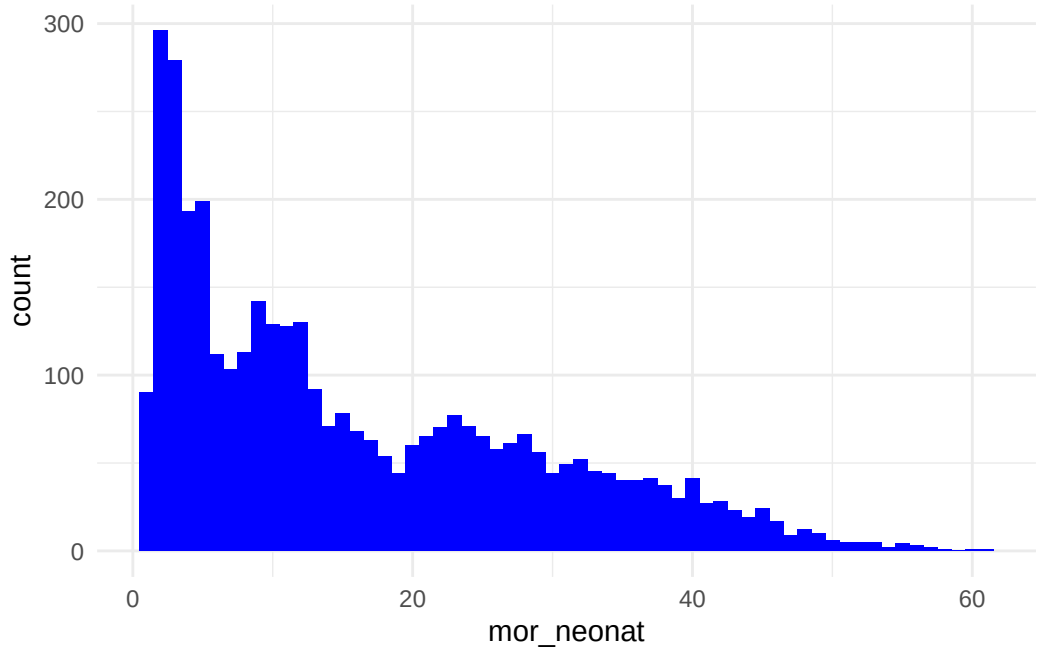




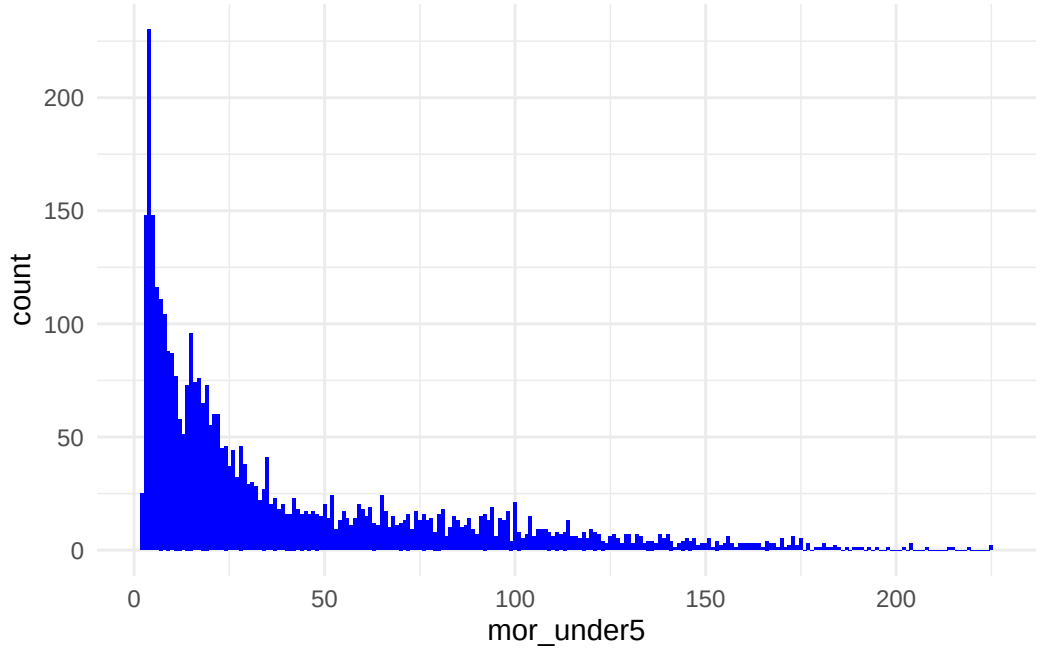

```
Warning: Removed 20 rows containing non-finite values (`stat_bin()`).
```



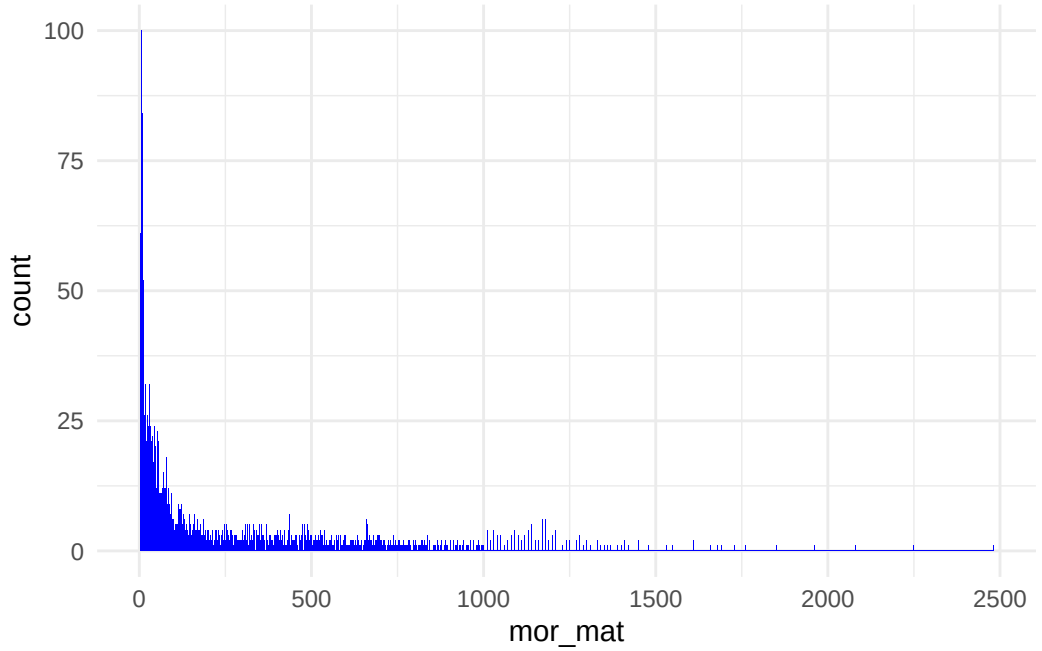

```
Warning: Removed 20 rows containing non-finite values (`stat_bin()`).
```




```
Warning: Removed 20 rows containing non-finite values (`stat_bin()`).
```




```
Warning: Removed 426 rows containing non-finite values (`stat_bin()`).
```



2730 (73%) rows of data had 0 battle-related deaths in the study time-frame.

```
totdeath_counts <- eda_data %>%
  group_by(totdeath) %>%
  summarize(count = n())

print(totdeath_counts)
```

```
# A tibble: 533 x 2
   totdeath count
      <int> <int>
 1        0  2730
 2        1    57
 3        2    32
 4        3    21
 5        4    22
 6        5    16
 7        6    14
 8        7     9
 9        8    13
10        9     8
```

```
# i 523 more rows
```

Summary of deaths by Country.

- Syria had the largest number of battle-related deaths

- Sierra Leone had the most maternal, infant, and under 5 deaths

- Pakistan had the most neonatal deaths

- Andora, Dominica, and Marshall Islands had no maternal death data

- Puerto Rico had no infant, neonatal, or under5 data. We may wish to exclude Puerto Rico from subsequent analyses.

```
# Total deaths table with NA count by country
eda_data %>%
  group_by(country_name) %>%
  summarise(
    total_deaths = sum(totdeath, na.rm = TRUE),
    total_deaths_NA = sum(is.na(totdeath))
  ) %>%
  arrange(desc(total_deaths))
```

```
# A tibble: 186 x 3
   country_name                    total_deaths total_deaths_NA
   <chr>                                  <int>           <int>
 1 Syria                                 386891               0
 2 Afghanistan                           171391               0
 3 Iraq                                   91429               0
 4 Ethiopia                               87066               0
 5 Democratic Republic of the Congo       52492               0
 6 Sudan                                  51355               0
 7 Nigeria                                51114               0
 8 Pakistan                               40789               0
 9 India                                  32704               0
10 Mexico                                 32686               0
# i 176 more rows
```

```r
# Maternal deaths table with NA count by country
eda_data %>%
  group_by(country_name) %>%
  summarise(
    maternal_deaths = sum(mor_mat, na.rm = TRUE),
    maternal_deaths_NA = sum(is.na(mor_mat))
  ) %>%
  arrange(desc(maternal_deaths))
```

```
# A tibble: 186 x 3
   country_name             maternal_deaths maternal_deaths_NA
   <chr>                              <int>              <int>
 1 Sierra Leone                       28280                  2
 2 South Sudan                        23720                  2
 3 Chad                               22920                  2
 4 Central African Republic           19536                  2
 5 Nigeria                            18573                  2
 6 Afghanistan                        18258                  2
 7 Somalia                            18017                  2
 8 Guinea-Bissau                      15840                  2
 9 Guinea                             14940                  2
10 Mauritania                         14666                  2
# i 176 more rows
```

```r
# Infant deaths table with NA count by country
eda_data %>%
  group_by(country_name) %>%
  summarise(
    infant_deaths = sum(mor_inf, na.rm = TRUE),
    infant_deaths_NA = sum(is.na(mor_inf))
  ) %>%
  arrange(desc(infant_deaths))
```

```
# A tibble: 186 x 3
   country_name                    infant_deaths infant_deaths_NA
   <chr>                                   <dbl>            <int>
 1 Sierra Leone                            2196.                0
 2 Central African Republic                1952.                0
 3 Somalia                                 1884.                0
 4 Nigeria                                 1764.                0
 5 Democratic Republic of the Congo        1705.                0
```

```
 6 Chad                                  1695.                0
 7 Equatorial Guinea                     1636                 0
 8 Angola                                1631.                0
 9 Liberia                               1622.                0
10 Mali                                  1578                 0
# i 176 more rows
```

```r
# Neonatal deaths table with NA count by country
eda_data %>%
  group_by(country_name) %>%
  summarise(
    neonatal_deaths = sum(mor_neonat, na.rm = TRUE),
    neonatal_deaths_NA = sum(is.na(mor_neonat))
  ) %>%
  arrange(desc(neonatal_deaths))
```

```
# A tibble: 186 x 3
   country_name             neonatal_deaths neonatal_deaths_NA
   <chr>                              <dbl>              <int>
 1 Pakistan                            989.                  0
 2 Afghanistan                         965.                  0
 3 South Sudan                         911.                  0
 4 Guinea-Bissau                       910.                  0
 5 Central African Republic            910.                  0
 6 Somalia                             858.                  0
 7 Lesotho                             845.                  0
 8 Sierra Leone                        805                   0
 9 Mali                                801.                  0
10 Cote d'Ivoire                       795                   0
# i 176 more rows
```

```r
# Under-5 deaths table with NA count by country
eda_data %>%
  group_by(country_name) %>%
  summarise(
    under5_deaths = sum(mor_under5, na.rm = TRUE),
    under5_deaths_NA = sum(is.na(mor_under5))
  ) %>%
  arrange(desc(under5_deaths))
```

```
# A tibble: 186 x 3
```

```
   country_name              under5_deaths under5_deaths_NA
   <chr>                             <dbl>            <int>
 1 Sierra Leone                      3334.                0
 2 Somalia                           3077                 0
 3 Chad                              2988.                0
 4 Central African Republic          2861.                0
 5 Nigeria                           2856                 0
 6 Niger                             2747.                0
 7 Mali                              2713.                0
 8 Angola                            2623.                0
 9 Burkina Faso                      2582.                0
10 Guinea                            2471                 0
# i 176 more rows
```

Summary of deaths by year

- The most battle-related deaths occurred in 2015

- The most maternal, infant, neonatal, and under 5 deaths occurred in 2000

- There is no maternal death data in 2018/2019

```r
# Total deaths by year with NA count
eda_data %>%
  group_by(year) %>%
  summarise(
    total_deaths = sum(totdeath, na.rm = TRUE),
    total_deaths_NA = sum(is.na(totdeath))
  ) %>%
  arrange(desc(total_deaths))
```

```
# A tibble: 20 x 3
    year total_deaths total_deaths_NA
   <int>        <int>           <int>
 1  2015       146657               0
 2  2016       127725               0
 3  2017       112422               0
 4  2014       111309               0
 5  2018       104701               0
 6  2000        98808               0
```

```
 7  2001         93703                0
 8  2013         86099                0
 9  2019         86029                0
10  2010         47424                0
11  2003         39027                0
12  2012         38964                0
13  2009         37231                0
14  2004         36931                0
15  2002         36922                0
16  2005         33699                0
17  2011         31123                0
18  2008         28153                0
19  2007         26917                0
20  2006         19522                0
```

```r
# Maternal deaths by year with NA count
eda_data %>%
  group_by(year) %>%
  summarise(
    maternal_deaths = sum(mor_mat, na.rm = TRUE),
    maternal_deaths_NA = sum(is.na(mor_mat))
  ) %>%
  arrange(desc(maternal_deaths))
```

```
# A tibble: 20 x 3
    year maternal_deaths maternal_deaths_NA
   <int>           <int>              <int>
 1  2000           50686                  3
 2  2001           49172                  3
 3  2002           47919                  3
 4  2003           46334                  3
 5  2004           44579                  3
 6  2005           42522                  3
 7  2006           40770                  3
 8  2007           39156                  3
 9  2008           37822                  3
10  2009           36611                  3
11  2010           35357                  3
12  2011           34245                  3
13  2012           33319                  3
14  2013           32541                  3
15  2014           31940                  3
```

```
16  2015           31083                  3
17  2016           30226                  3
18  2017           29493                  3
19  2018               0                186
20  2019               0                186
```

```r
# Infant deaths by year with NA count
eda_data %>%
  group_by(year) %>%
  summarise(
    infant_deaths = sum(mor_inf, na.rm = TRUE),
    infant_deaths_NA = sum(is.na(mor_inf))
  ) %>%
  arrange(desc(infant_deaths))
```

```
# A tibble: 20 x 3
    year infant_deaths infant_deaths_NA
   <int>         <dbl>            <int>
 1  2000         7257                 1
 2  2001         7000.                1
 3  2002         6747.                1
 4  2003         6505.                1
 5  2004         6280.                1
 6  2005         6044.                1
 7  2006         5827                 1
 8  2007         5632.                1
 9  2008         5461.                1
10  2009         5272.                1
11  2010         5135.                1
12  2011         4931.                1
13  2012         4787.                1
14  2013         4650.                1
15  2014         4525.                1
16  2015         4401                 1
17  2016         4285                 1
18  2017         4167.                1
19  2018         4054.                1
20  2019         3953.                1
```

```r
# Neonatal deaths by year with NA count
eda_data %>%
  group_by(year) %>%
```

```
  summarise(
    neonatal_deaths = sum(mor_neonat, na.rm = TRUE),
    neonatal_deaths_NA = sum(is.na(mor_neonat))
  ) %>%
  arrange(desc(neonatal_deaths))
```

```
# A tibble: 20 x 3
    year neonatal_deaths neonatal_deaths_NA
   <int>           <dbl>              <int>
 1  2000           3772.                  1
 2  2001           3668.                  1
 3  2002           3567                   1
 4  2003           3471.                  1
 5  2004           3379.                  1
 6  2005           3292.                  1
 7  2006           3210.                  1
 8  2007           3129.                  1
 9  2008           3052.                  1
10  2009           2978.                  1
11  2010           2908.                  1
12  2011           2838.                  1
13  2012           2776.                  1
14  2013           2713.                  1
15  2014           2654.                  1
16  2015           2599.                  1
17  2016           2544.                  1
18  2017           2490.                  1
19  2018           2435.                  1
20  2019           2382.                  1
```

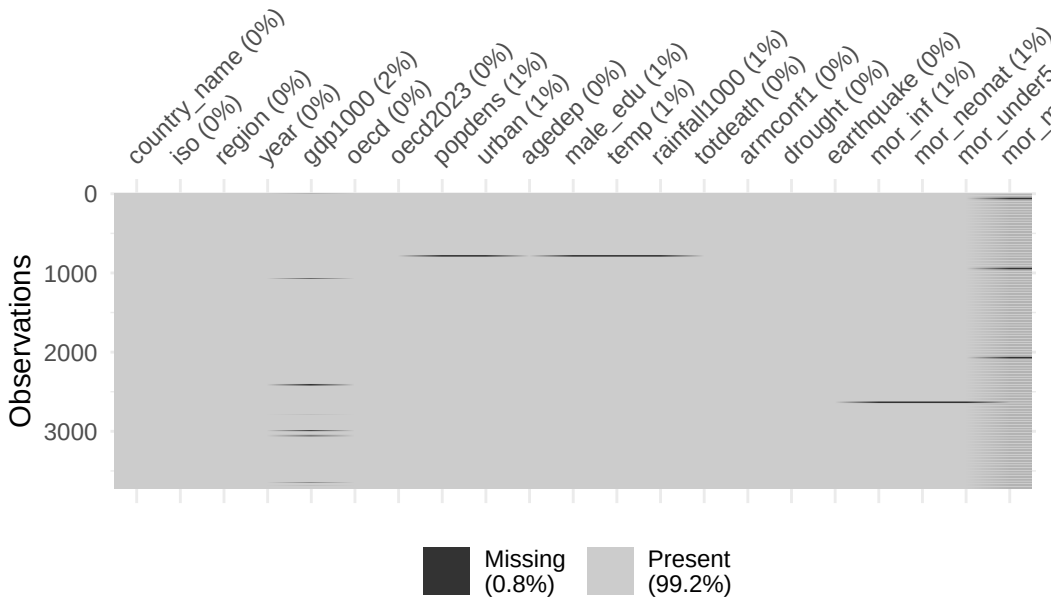
```
# Under-5 deaths by year with NA count
eda_data %>%
  group_by(year) %>%
  summarise(
    under5_deaths = sum(mor_under5, na.rm = TRUE),
    under5_deaths_NA = sum(is.na(mor_under5))
  ) %>%
  arrange(desc(under5_deaths))
```

```
# A tibble: 20 x 3
    year under5_deaths under5_deaths_NA
```

	<int>	<dbl>	<int>
1	2000	10619	1
2	2001	10212.	1
3	2002	9805.	1
4	2003	9413.	1
5	2004	9059.	1
6	2005	8672.	1
7	2006	8318.	1
8	2007	7983.	1
9	2008	7695.	1
10	2009	7363.	1
11	2010	7171	1
12	2011	6777.	1
13	2012	6538.	1
14	2013	6315.	1
15	2014	6118.	1
16	2015	5925.	1
17	2016	5727.	1
18	2017	5541.	1
19	2018	5373.	1
20	2019	5224.	1

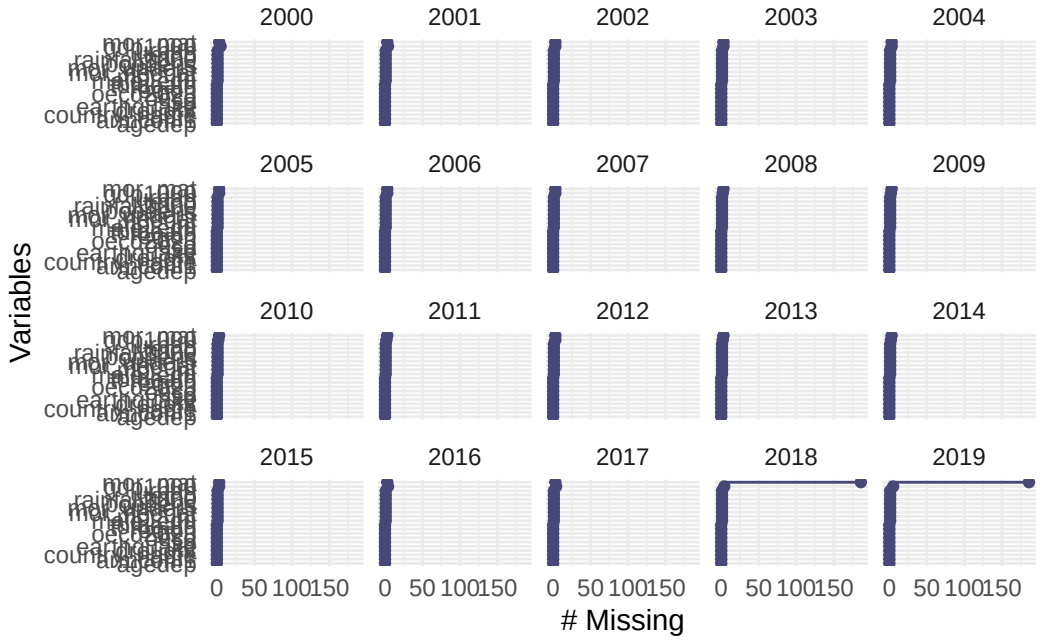
Visualizing the amount of missing data across all variables. There is a high level of missingness for maternal mortality (11% missing). All other variables have less than 5% missing data.

```
vis_miss(eda_data)
```



Further visualizing the missingness by year. 2018/2019 has the most missing data for maternal mortality.

```
gg_miss_var(eda_data, facet = year)
```



Checking the correlations of the numeric variables

```
numeric_data <- eda_data[, sapply(eda_data, is.numeric)]

numeric_data <- numeric_data %>%
  select(-year)

cor(numeric_data, use = "complete.obs")
```

```
                gdp1000     popdens       urban     agedep    male_edu
gdp1000      1.00000000  0.27790004  0.35183092 -0.4591014  0.57894538
popdens      0.27790004  1.00000000  0.73026085 -0.3750241  0.30142268
urban        0.35183092  0.73026085  1.00000000 -0.3692659  0.36934543
agedep      -0.45910137 -0.37502408 -0.36926592  1.0000000 -0.68724808
male_edu     0.57894538  0.30142268  0.36934543 -0.6872481  1.00000000
temp        -0.37911827 -0.11110113 -0.05153235  0.4115461 -0.66002036
rainfall1000 -0.13121810 -0.07238190 -0.04762704  0.1094754 -0.16190839
totdeath    -0.06476673 -0.02690464  0.01194689  0.1166521 -0.04610287
mor_inf     -0.49826673 -0.34580395 -0.36724226  0.8097231 -0.73815037
mor_neonat  -0.54118022 -0.34206796 -0.36827308  0.7892814 -0.76629253
mor_under5  -0.45486705 -0.34168813 -0.34948188  0.8156673 -0.72335088
```

```
mor_mat      -0.37274140 -0.30251831 -0.30624597  0.7651519 -0.67078776
                    temp rainfall1000    totdeath     mor_inf  mor_neonat
gdp1000      -0.37911827  -0.13121810 -0.06476673 -0.49826673 -0.54118022
popdens      -0.11110113  -0.07238190 -0.02690464 -0.34580395 -0.34206796
urban        -0.05153235  -0.04762704  0.01194689 -0.36724226 -0.36827308
agedep        0.41154610   0.10947539  0.11665213  0.80972314  0.78928144
male_edu     -0.66002036  -0.16190839 -0.04610287 -0.73815037 -0.76629253
temp          1.00000000   0.40678877  0.01385887  0.46634829  0.48137371
rainfall1000  0.40678877   1.00000000 -0.06472609  0.08515842  0.06039748
totdeath      0.01385887  -0.06472609  1.00000000  0.07730430  0.08558692
mor_inf       0.46634829   0.08515842  0.07730430  1.00000000  0.95771780
mor_neonat    0.48137371   0.06039748  0.08558692  0.95771780  1.00000000
mor_under5    0.45374010   0.05868111  0.07648536  0.98538338  0.92511366
mor_mat       0.41551273   0.08338977  0.07601503  0.87675388  0.82742960
              mor_under5     mor_mat
gdp1000      -0.45486705 -0.37274140
popdens      -0.34168813 -0.30251831
urban        -0.34948188 -0.30624597
agedep        0.81566729  0.76515191
male_edu     -0.72335088 -0.67078776
temp          0.45374010  0.41551273
rainfall1000  0.05868111  0.08338977
totdeath      0.07648536  0.07601503
mor_inf       0.98538338  0.87675388
mor_neonat    0.92511366  0.82742960
mor_under5    1.00000000  0.89870312
mor_mat       0.89870312  1.00000000
```

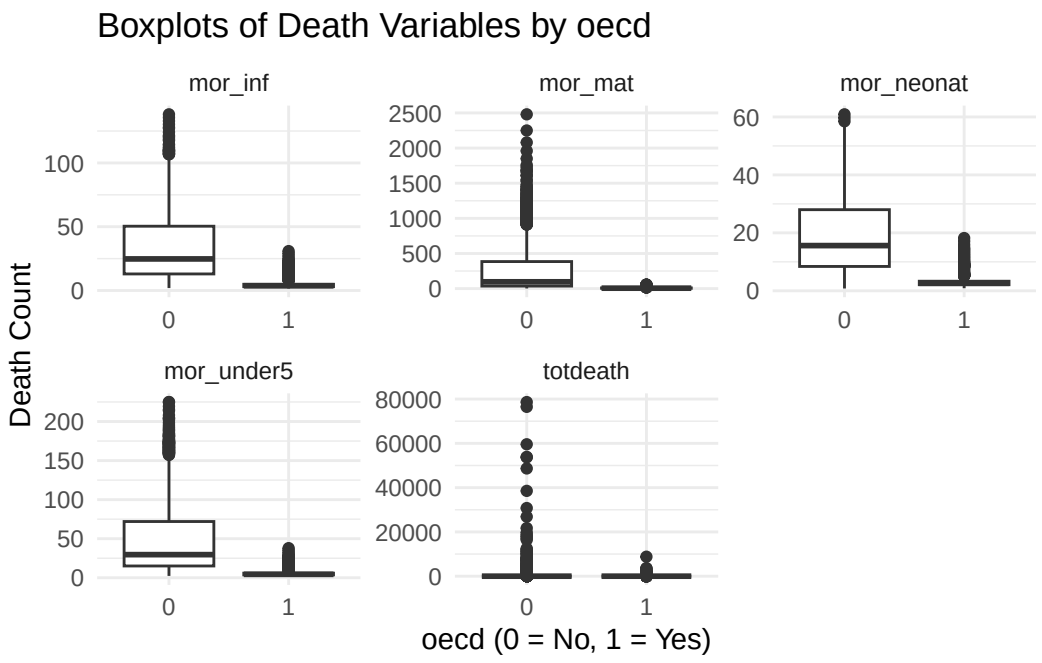
Generating boxplots to show the relationship between the mortality variables and the binary variables.

- OECD countries have far fewer battle-related deaths as well as maternal, neonatal, infant, and under 5 deaths
- Median death count for all mortality variables is higher when there is an armed conflict
- Median death count for all mortality variables is only slightly higher when there is an earthquake
- Median death count for all mortality variables is higher when there is a drought

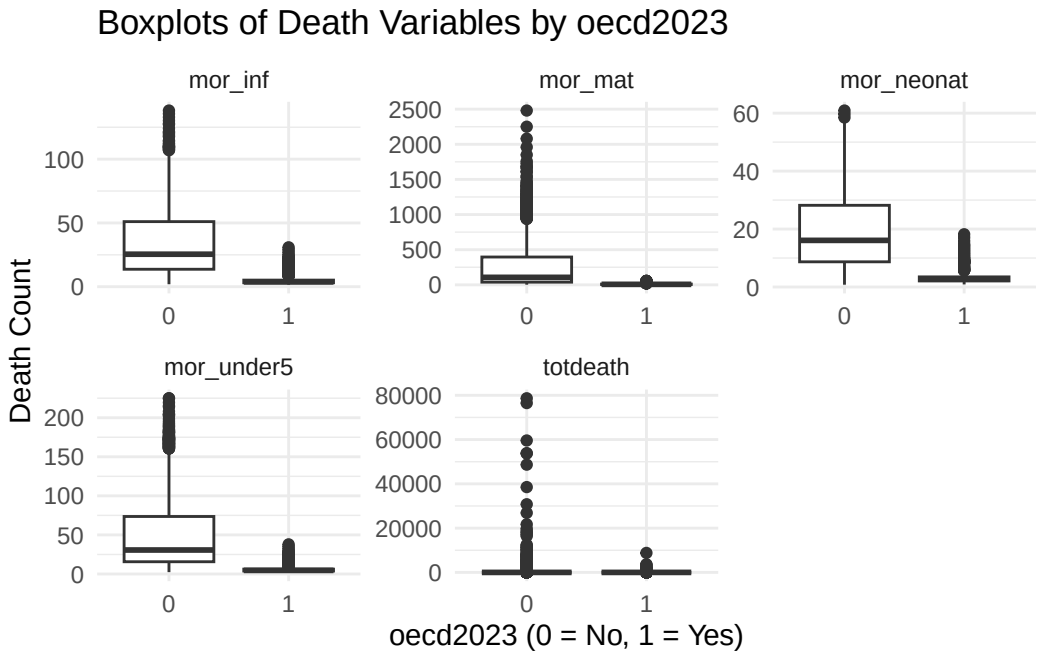
```r
generate_boxplots <- function(data, binary_vars) {
  # Reshape the data to long format for the death variables
  data_long <- data %>%
    pivot_longer(cols = c(mor_inf, mor_mat, mor_under5, mor_neonat, totdeath),
                 names_to = "death_type", values_to = "death_value")

  # Loop through each binary variable
  for (binary_var in binary_vars) {
    # Create boxplot for the current binary variable
    p <- ggplot(data_long, aes_string(x = binary_var, y = "death_value"))+
      geom_boxplot() +
      facet_wrap(~ death_type, scales = "free") +
      labs(x = paste(binary_var, "(0 = No, 1 = Yes)"), y = "Death Count",
           title = paste("Boxplots of Death Variables by", binary_var)) +
      theme_minimal()

    # Print the plot
    print(p)
  }
}


binary_variables <- c("oecd", "oecd2023", "armconf1", "earthquake", "drought")

generate_boxplots(eda_data, binary_variables)
```

```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```

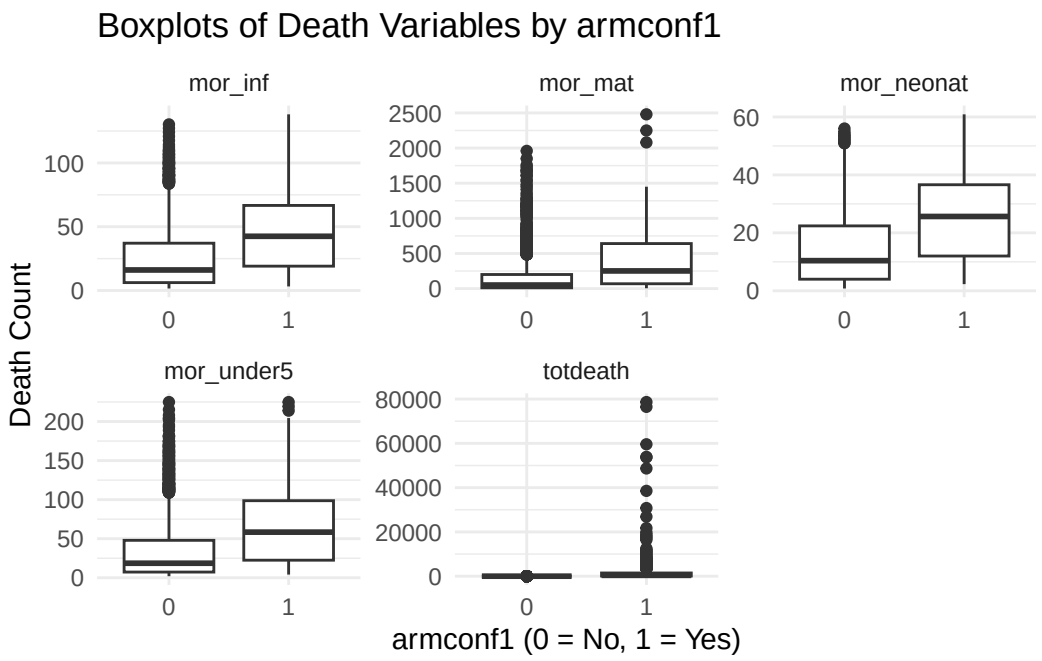
```
Warning: Removed 486 rows containing non-finite values (`stat_boxplot()`).
```



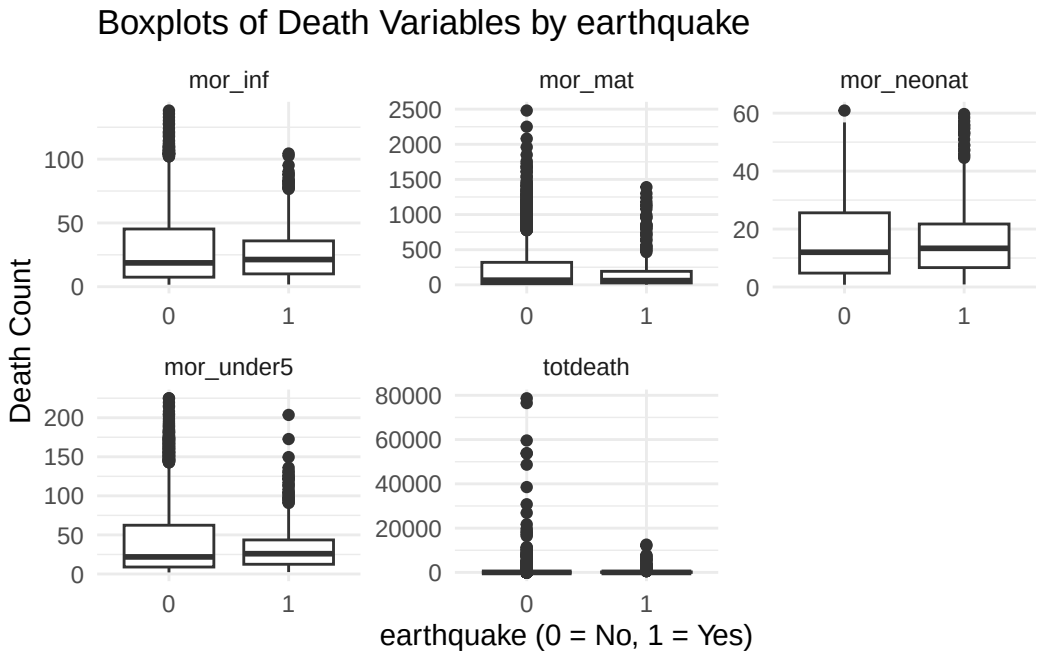

```
Warning: Removed 486 rows containing non-finite values (`stat_boxplot()`).
```



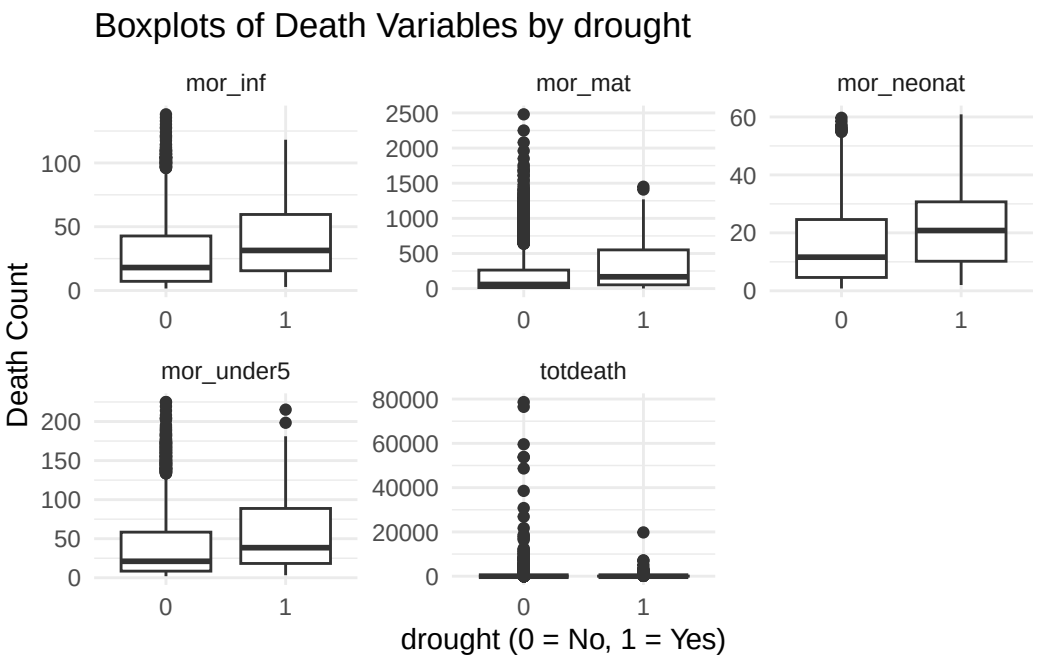

```
Warning: Removed 486 rows containing non-finite values (`stat_boxplot()`).
```




```
Warning: Removed 486 rows containing non-finite values (`stat_boxplot()`).
```




```
Warning: Removed 486 rows containing non-finite values (`stat_boxplot()`).
```



Creating scatterplots to visualize the relationship between the mortality variables and the other continuous variables.

- There is a negative relationship between all mortality variables and population density, urban residence, and male education
- There is a positive relationship between the mortality variables and age dependency and temperature
- The relationship between the mortality variables and rainfall is less clear and warrants further exploration

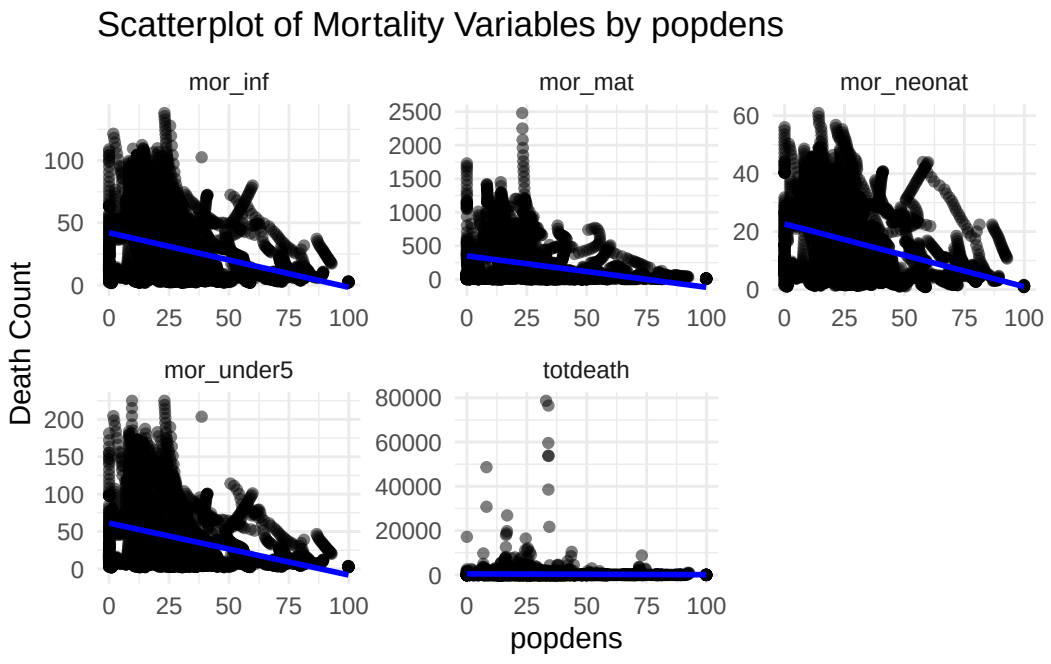
```
generate_scatterplots <- function(data, continuous_vars) {
  # Reshape the data to long format for the death variables
  data_long <- data %>%
    pivot_longer(cols = c(mor_inf, mor_mat, mor_under5, mor_neonat, totdeath),
                 names_to = "death_type", values_to = "death_value")

  # Loop through each binary variable
  for (continuous_vars in continuous_vars) {
    # Create boxplot for the current binary variable
    p <- ggplot(data_long, aes_string(x = continuous_vars, y = "death_value"))+
      geom_point(alpha=.5) +
```

```
      geom_smooth(method = "lm", se = FALSE, color = "blue") +
      facet_wrap(~ death_type, scales = "free") +
      labs(x = paste(continuous_vars), y = "Death Count",
           title = paste("Scatterplot of Mortality Variables by", continuous_vars)) +
      theme_minimal()

    # Print the plot
    print(p)
  }
}


continuous_vars <- c("popdens", "urban", "agedep", "male_edu", "temp", "rainfall1000")

generate_scatterplots(eda_data, continuous_vars)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

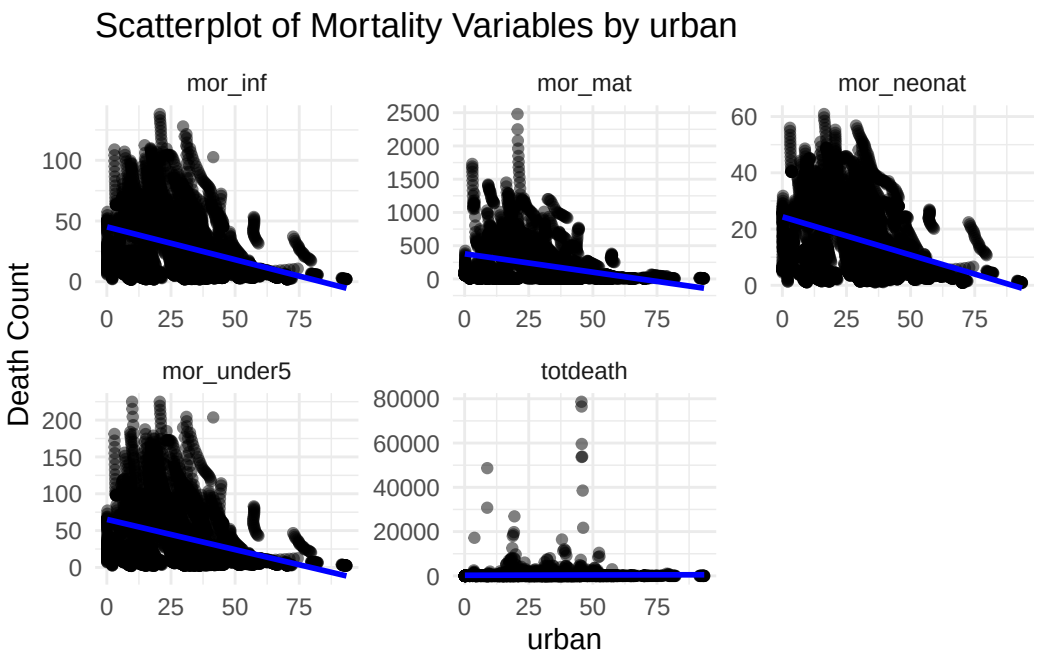
```
Warning: Removed 584 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 584 rows containing missing values (`geom_point()`).
```



```
`geom_smooth()` using formula = 'y ~ x'
```

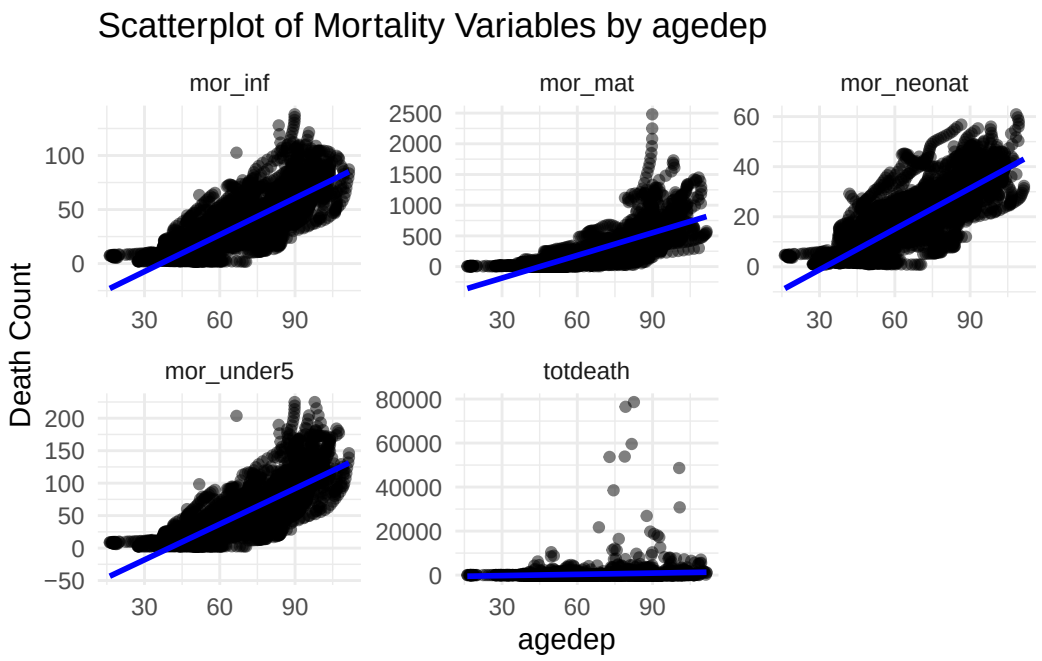
```
Warning: Removed 584 rows containing non-finite values (`stat_smooth()`).
Removed 584 rows containing missing values (`geom_point()`).
```




```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 486 rows containing non-finite values (`stat_smooth()`).
```

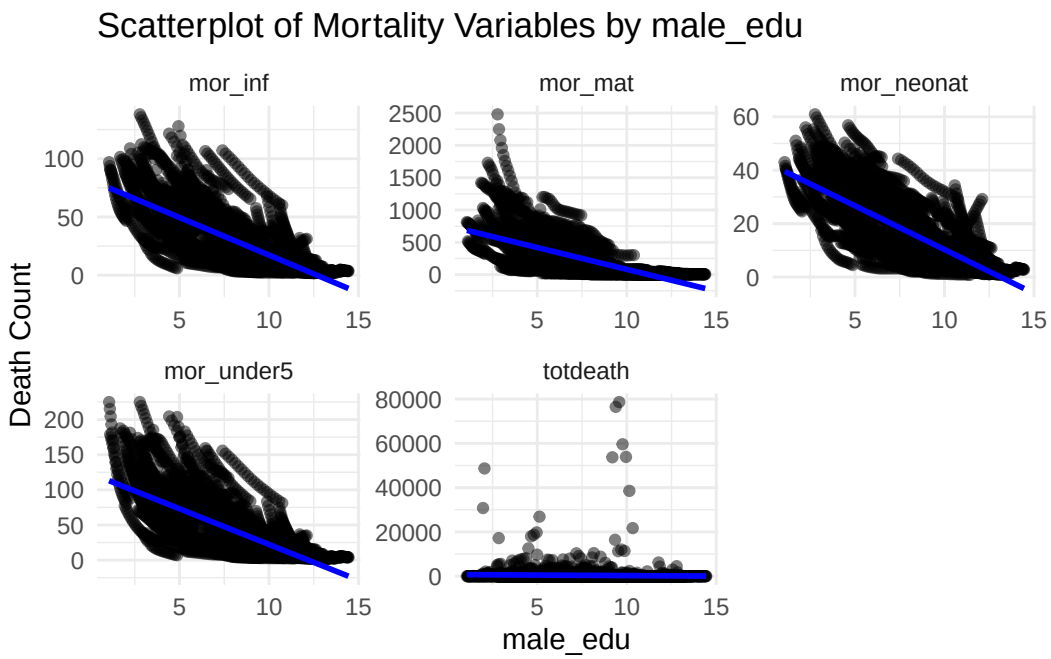
```
Warning: Removed 486 rows containing missing values (`geom_point()`).
```




```
`geom_smooth()` using formula = 'y ~ x'
```

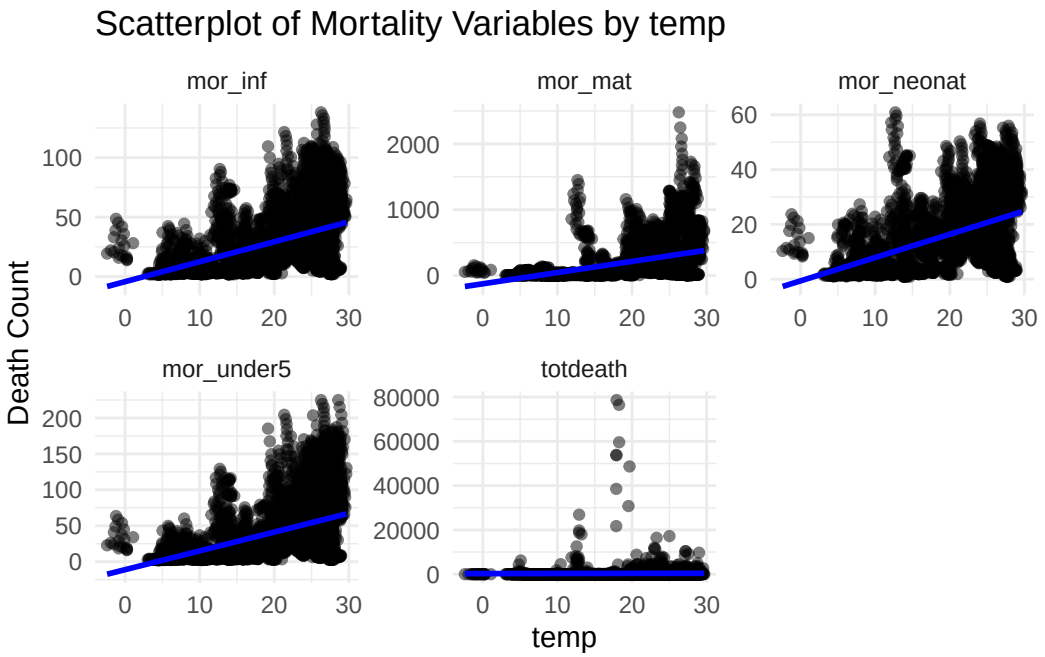
```
Warning: Removed 584 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 584 rows containing missing values (`geom_point()`).
```



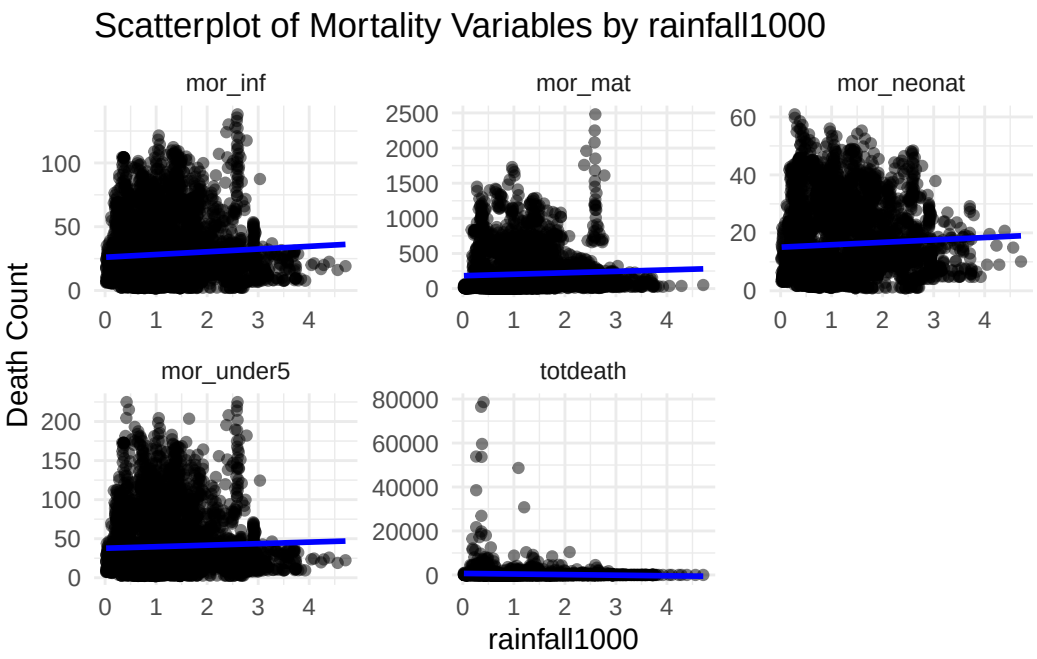

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 584 rows containing non-finite values (`stat_smooth()`).
Removed 584 rows containing missing values (`geom_point()`).
```




```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 584 rows containing non-finite values (`stat_smooth()`).
Removed 584 rows containing missing values (`geom_point()`).
```


Scatterplot of Mortality Variables by rainfall1000
mor_inf
mor_mat
mor_neonat
mor_under5
totdeath
Death Count
rainfall1000