EDA

Szewei Wang

2024-10-07

Data Loading and Initial Overview

```
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(naniar)
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.94 loaded
```

```
library(ggplot2)
library(dplyr)

# Load the dataset
data <- read.csv("/Users/wangsiwei/Desktop/version_control/armed_conflict/data/analytical/finaldata.csv

# Display the first few rows
head(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 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       1          0
## 2   59.7  125.2       0          1
## 3   58.5  121.1       0          1
## 4   57.2  116.9       0          1
## 5   55.9  112.6       0          1
## 6   54.6  108.4       0          1
```

```
# Overview of the structure of the dataset
str(data)
```

```
## '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 ...
##  $ totdeath    : int  5065 5394 5553 1157 944 817 1711 4982 7020 5660 ...
##  $ armconf1    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ matmor      : int  1450 1390 1300 1240 1180 1140 1120 1090 1030 993 ...
##  $ infmor      : num  90.5 87.9 85.3 82.7 80 77.3 74.6 71.9 69.2 66.7 ...
##  $ neomor      : num  60.9 59.7 58.5 57.2 55.9 54.6 53.2 51.7 50.3 48.9 ...
##  $ un5mor      : num  129 125 121 117 113 ...
##  $ 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 statistics
summary(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        totdeath          armconf1          matmor
##  Min.   :0.01993   Min.   :    0.0   Min.   :0.0000   Min.   :   2.0
##  1st Qu.:0.59146   1st Qu.:    0.0   1st Qu.:0.0000   1st Qu.:  17.0
##  Median :1.01288   Median :    0.0   Median :0.0000   Median :  66.0
##  Mean   :1.20216   Mean   :  361.1   Mean   :0.1892   Mean   : 210.6
##  3rd Qu.:1.68706   3rd Qu.:    2.0   3rd Qu.:0.0000   3rd Qu.: 299.8
##  Max.   :4.71081   Max.   :78644.0   Max.   :1.0000   Max.   :2480.0
##  NA's   :20                                          NA's   :426
##      infmor           neomor          un5mor          drought
##  Min.   :  1.60   Min.   : 0.80   Min.   :  2.00   Min.   :0.00000
##  1st Qu.:  7.60   1st Qu.: 4.90   1st Qu.:  9.00   1st Qu.:0.00000
##  Median : 18.90   Median :12.10   Median : 22.20   Median :0.00000
##  Mean   : 28.90   Mean   :16.18   Mean   : 40.50   Mean   :0.08737
##  3rd Qu.: 44.52   3rd Qu.:25.32   3rd Qu.: 61.33   3rd Qu.:0.00000
##  Max.   :138.10   Max.   :60.90   Max.   :224.90   Max.   :1.00000
##  NA's   :20       NA's   :20      NA's   :20
##    earthquake
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.08333
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

Insight: - The dataset contains various economic, demographic, and environmental indicators. - The structure overview shows numeric variables like `gdp1000`, `popdens`, and mortality rates. - The summary statistics reveal missing values in some columns (like GDP), which may need further attention.

```r
# Check for missing values
missing_values <- colSums(is.na(data))
missing_values
```

```
## country_name          ISO       region         year      gdp1000         OECD
##            0            0            0            0           62            0
##     OECD2023      popdens        urban       agedep     male_edu         temp
##            0           20           20            0           20           20
```

```
## rainfall1000     totdeath     armconf1       matmor       infmor       neomor
##           20            0            0          426           20           20
##       un5mor      drought   earthquake
##           20            0            0
```

Insight: - We can observe missing values in several columns, especially `gdp1000`. Handling these missing values will be essential for analysis and interpretation.

```
# Descriptive statistics for numeric variables
numeric_vars <- data %>% select(where(is.numeric))
summary(numeric_vars)
```

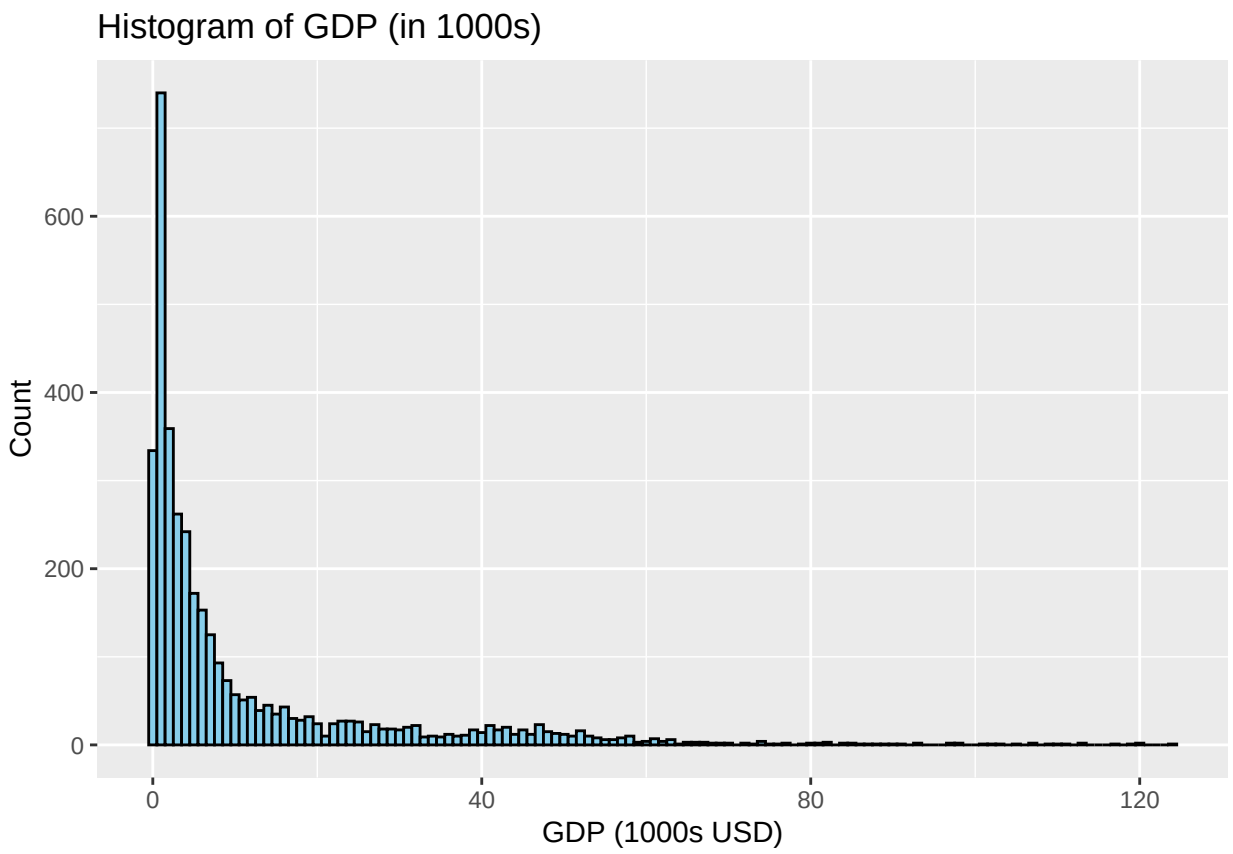
```
##       year         gdp1000             OECD          OECD2023
##  Min.   :2000   Min.   :  0.1105   Min.   :0.000   Min.   :0.0000
##  1st Qu.:2005   1st Qu.:  1.2383   1st Qu.:0.000   1st Qu.:0.0000
##  Median :2010   Median :  4.0719   Median :0.000   Median :0.0000
##  Mean   :2010   Mean   : 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   :0.08737   Mean   :0.08333
##  3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000
##
```

Insight: - The summary confirms wide variability across numeric variables. For example, `gdp1000` shows a high range of values, likely due to differences in economic output among countries.

Univariate Analysis

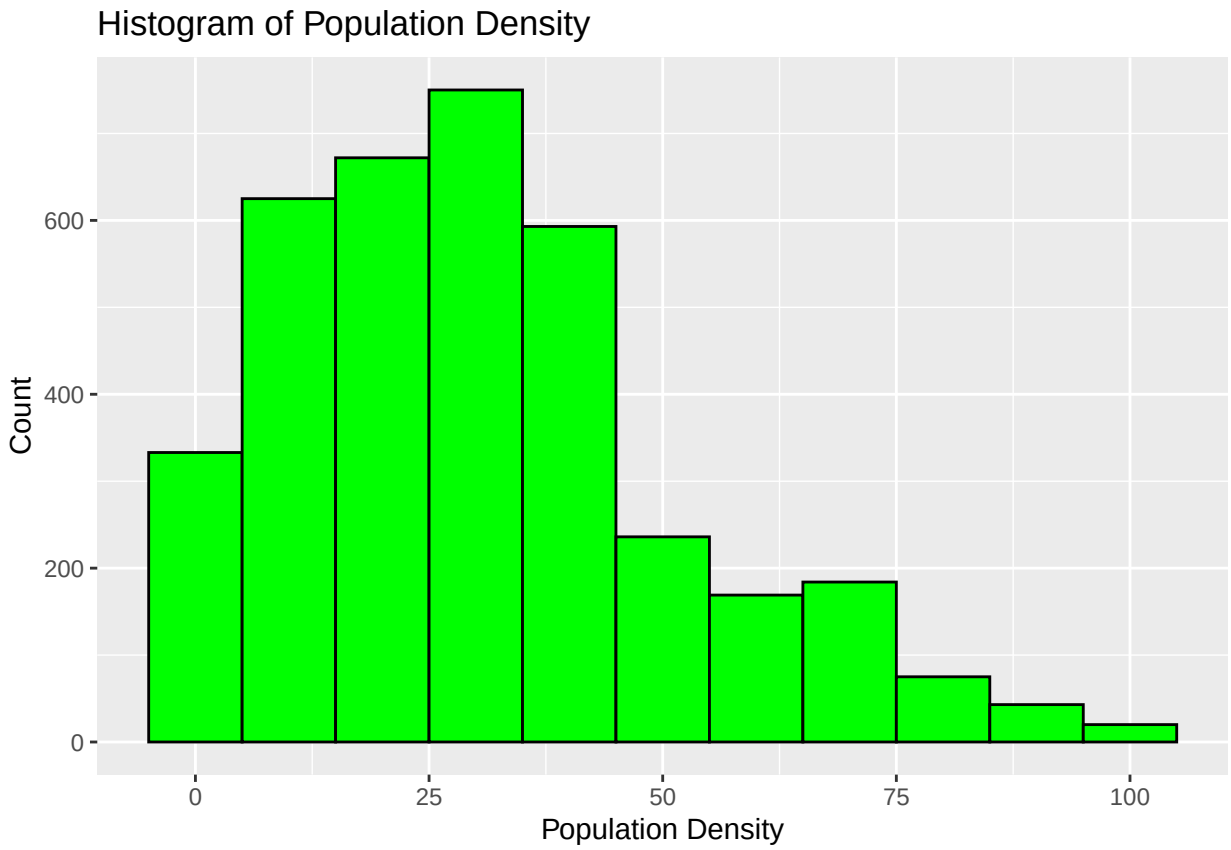
```
# GDP distribution
ggplot(data, aes(x = gdp1000)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Histogram of GDP (in 1000s)", x = "GDP (1000s USD)", y = "Count")
```

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




```
# Population density distribution
ggplot(data, aes(x = popdens)) +
  geom_histogram(binwidth = 10, fill = "green", color = "black") +
  labs(title = "Histogram of Population Density", x = "Population Density", y = "Count")
```

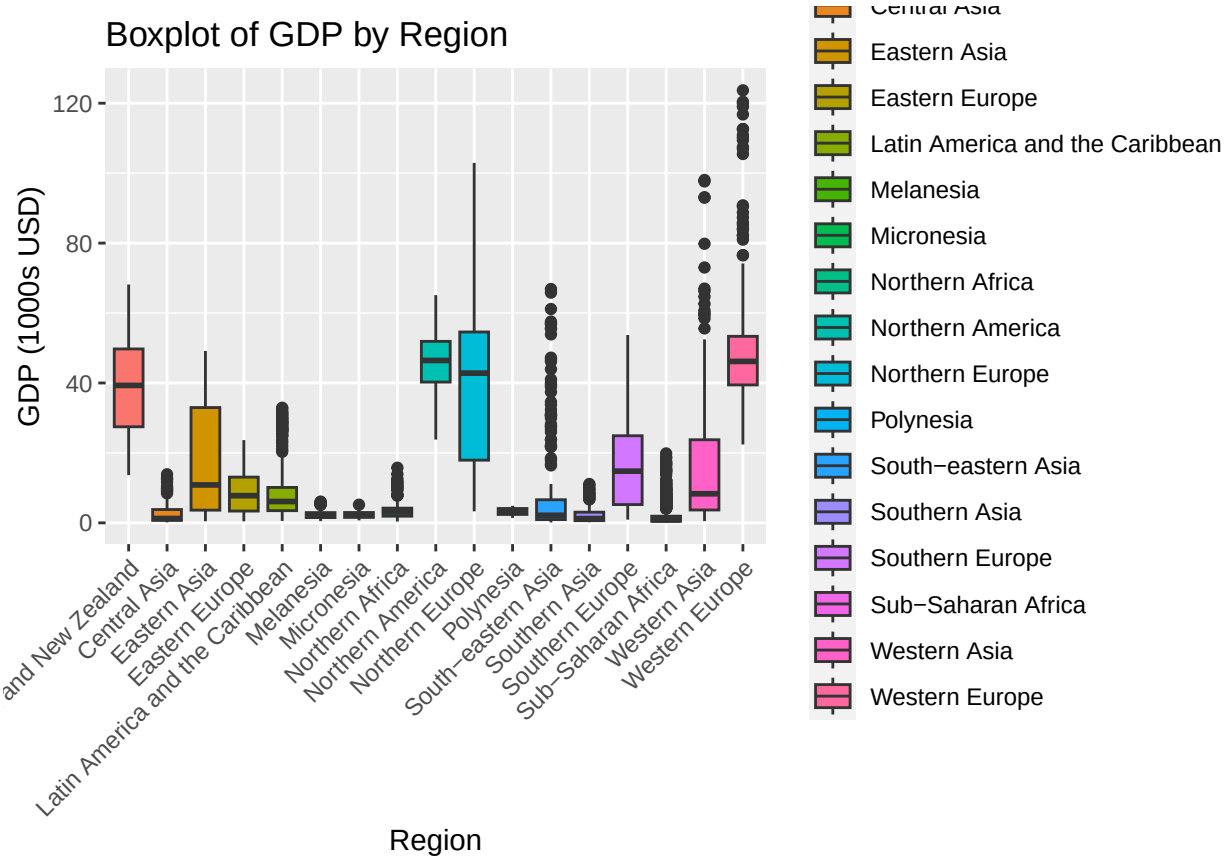
```
## Warning: Removed 20 rows containing non-finite values (‘stat_bin()‘).
```



Insight: - GDP distribution is right-skewed, suggesting that most countries have lower GDPs, while a few have significantly higher values. - Population density also shows a right-skew, with most countries having low-to-moderate densities and fewer countries having high densities.

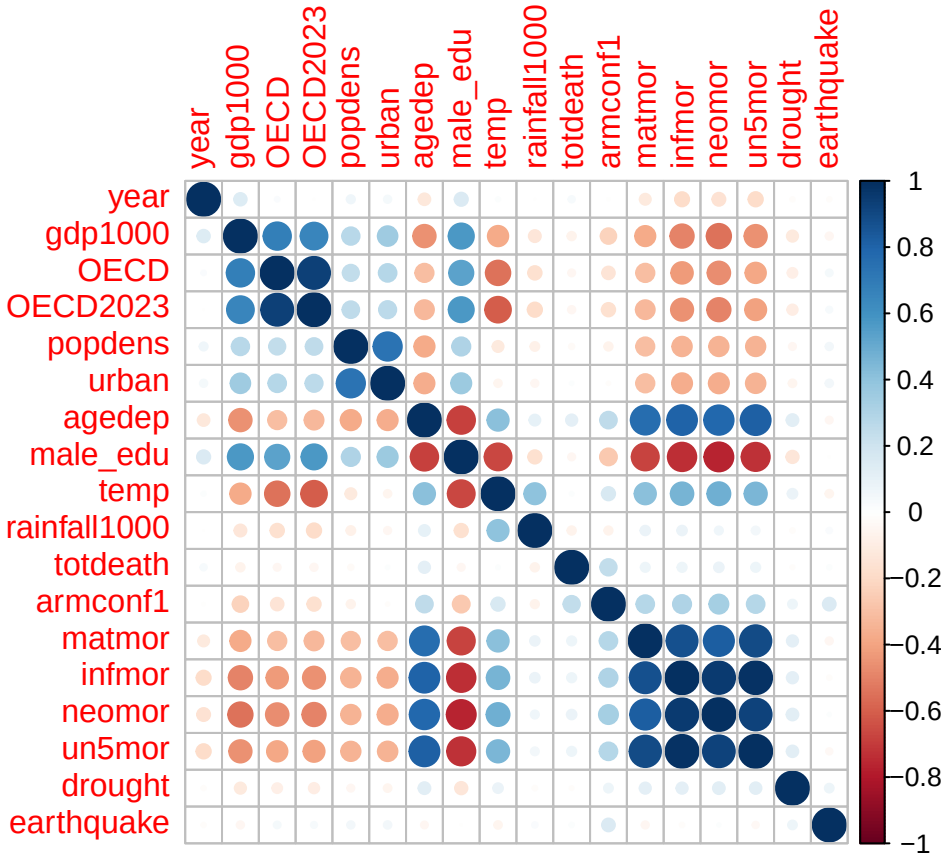
```
# Boxplot of GDP by region
ggplot(data, aes(x = region, y = gdp1000, fill = region)) +
  geom_boxplot() +
  labs(title = "Boxplot of GDP by Region", x = "Region", y = "GDP (1000s USD)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

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



Insight: - There are notable differences in GDP across regions, with some regions exhibiting significantly higher GDPs. The boxplot shows high variation within regions as well, indicating economic disparity.

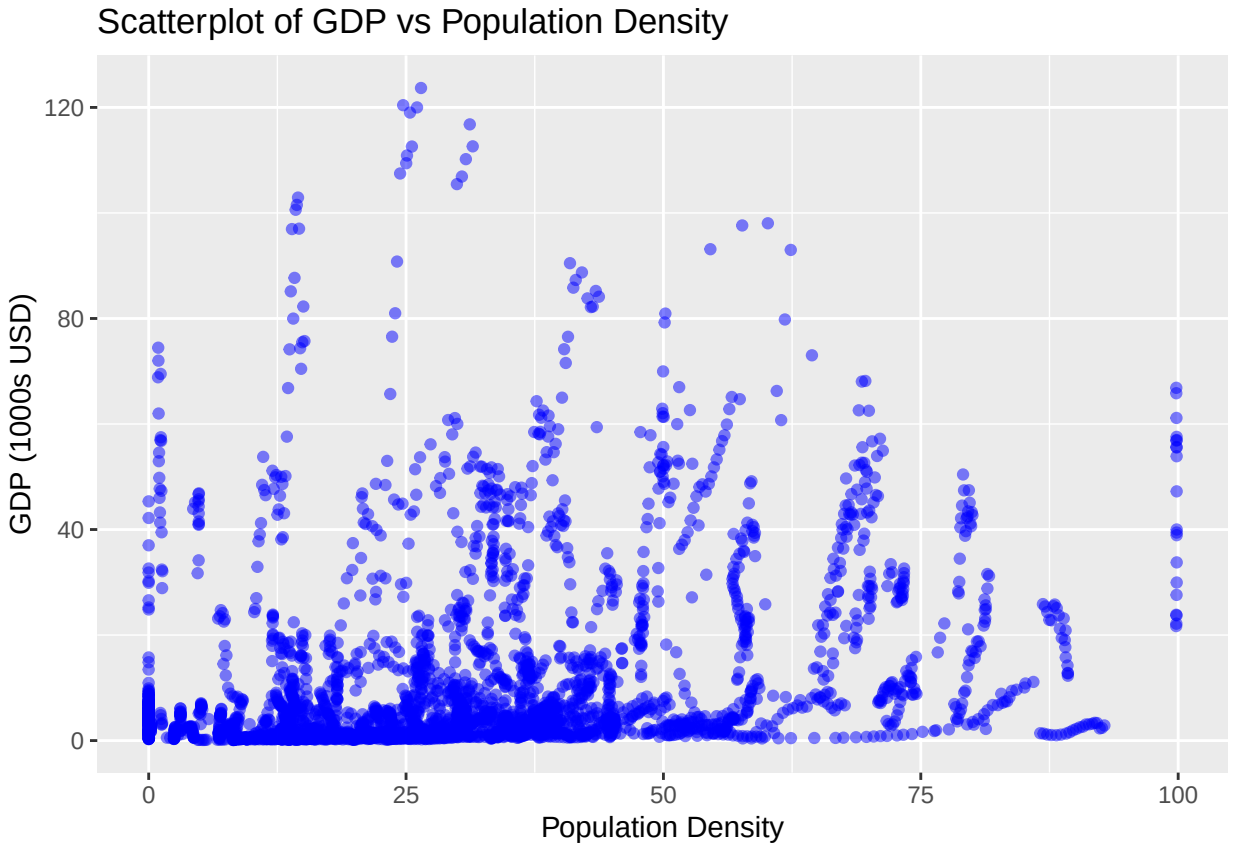
```
# Correlation matrix
corr_matrix <- cor(numeric_vars, use = "complete.obs")
corrplot(corr_matrix, method = "circle")
```



Insight: - Several numeric variables are highly correlated, such as mortality indicators (`infmor`, `neomor`, `matmor`), which makes sense as these reflect health outcomes.

```
# Scatterplot of GDP vs Population Density
ggplot(data, aes(x = popdens, y = gdp1000)) +
  geom_point(alpha = 0.5, color = "blue") +
  labs(title = "Scatterplot of GDP vs Population Density", x = "Population Density", y = "GDP (1000s US
```

```
## Warning: Removed 82 rows containing missing values (`geom_point()`).
```



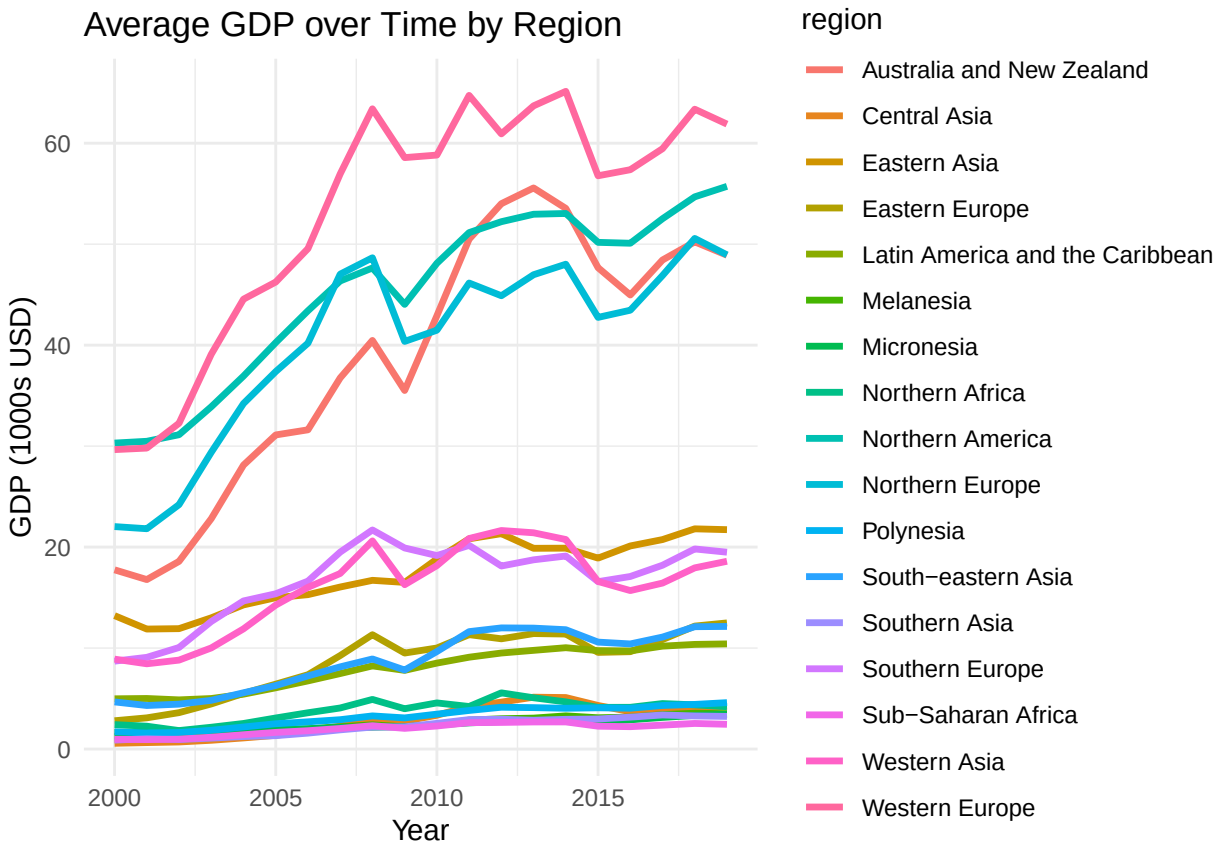
Insight: - There doesn't seem to be a strong linear relationship between population density and GDP. This suggests that while population density might influence GDP, other factors are more significant.

Regional Trends

```
# Regional GDP trends over time
ggplot(data, aes(x = year, y = gdp1000, color = region)) +
  geom_line(stat = "summary", fun = "mean", size = 1.2) +
  labs(title = "Average GDP over Time by Region", x = "Year", y = "GDP (1000s USD)") +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

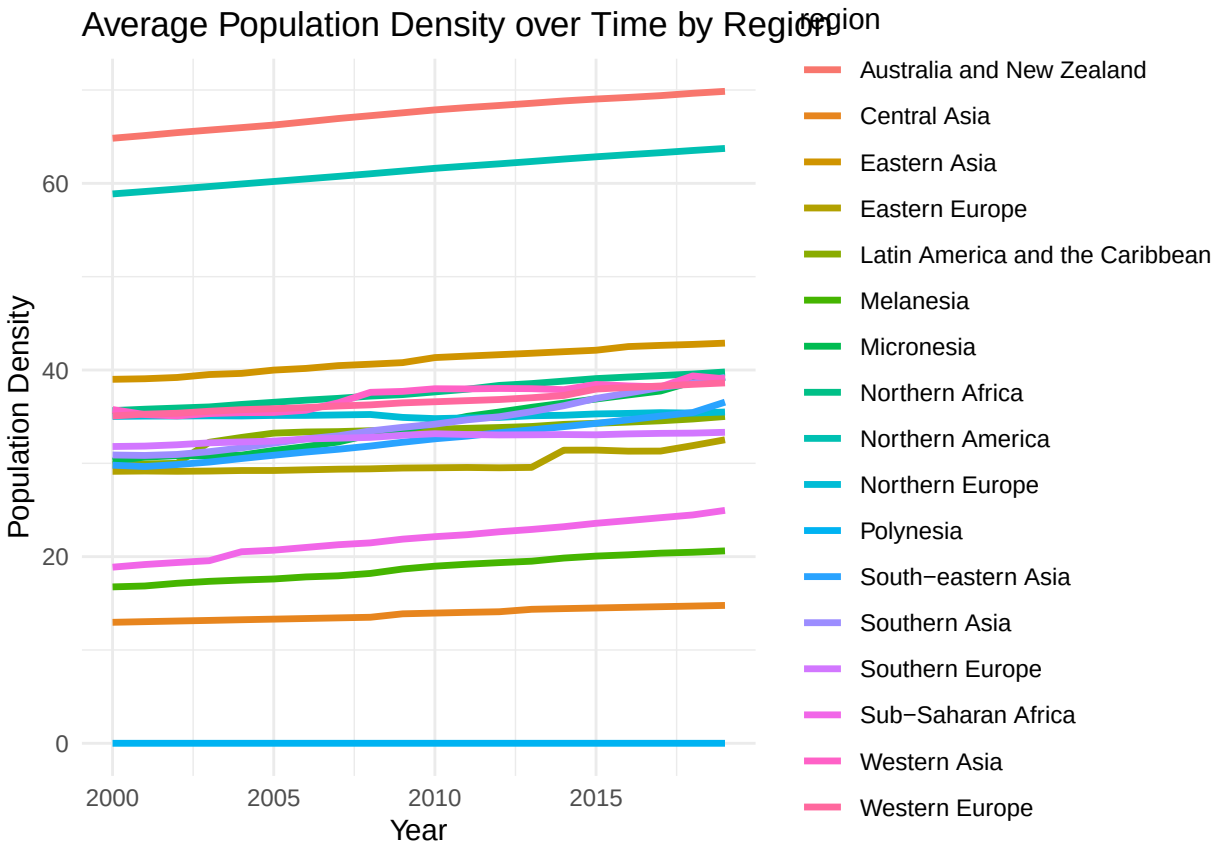
```
## Warning: Removed 62 rows containing non-finite values (‘stat_summary()‘).
```



Insight: - GDP has been increasing for some regions over time, though the rate of increase varies. Certain regions, like North America and Europe, exhibit higher GDPs throughout.

```
# Population density trends over time by region
ggplot(data, aes(x = year, y = popdens, color = region)) +
  geom_line(stat = "summary", fun = "mean", size = 1.2) +
  labs(title = "Average Population Density over Time by Region", x = "Year", y = "Population Density") +
  theme_minimal()
```

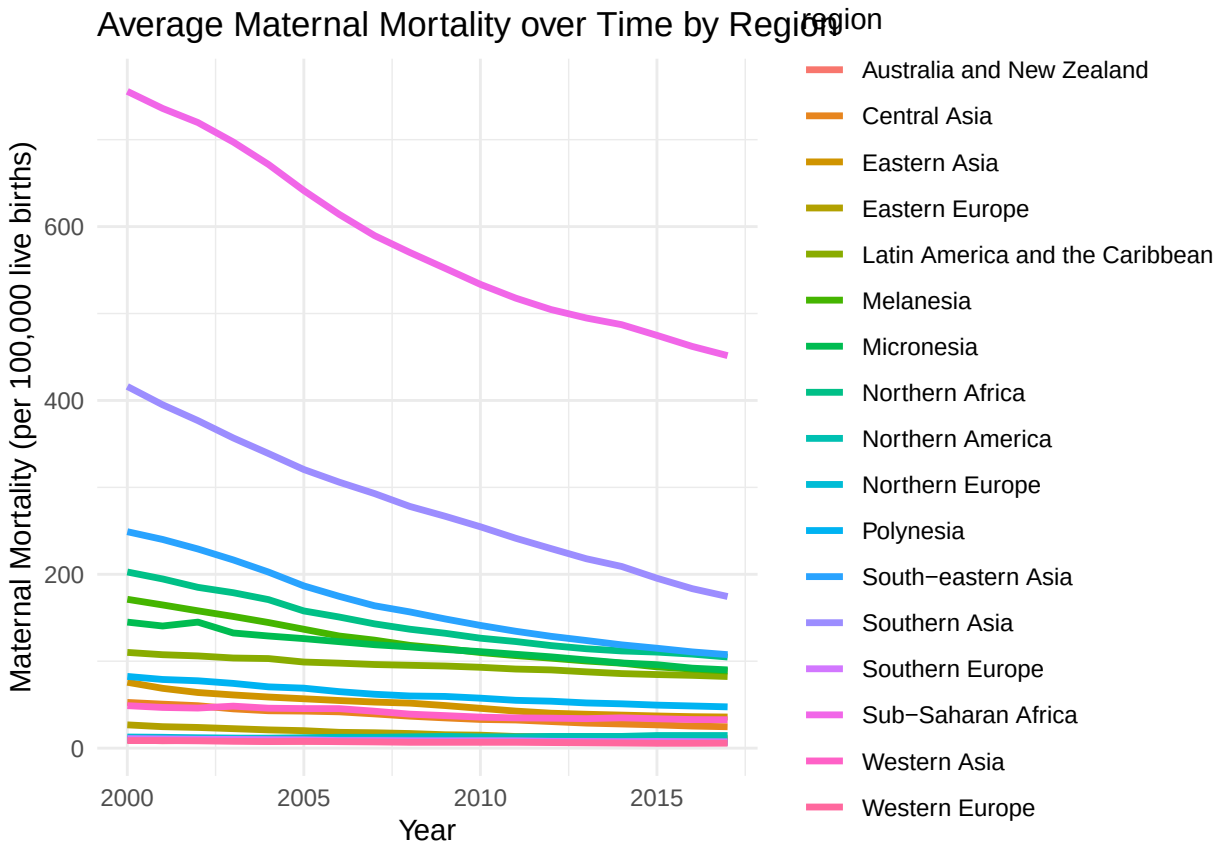
```
## Warning: Removed 20 rows containing non-finite values (`stat_summary()`).
```



Insight: - Population density appears relatively stable over time, with only slight variations. Some regions exhibit consistently higher population densities.

```
# Maternal mortality trends by region
ggplot(data, aes(x = year, y = matmor, color = region)) +
  geom_line(stat = "summary", fun = "mean", size = 1.2) +
  labs(title = "Average Maternal Mortality over Time by Region", x = "Year", y = "Maternal Mortality (p
  theme_minimal()
```

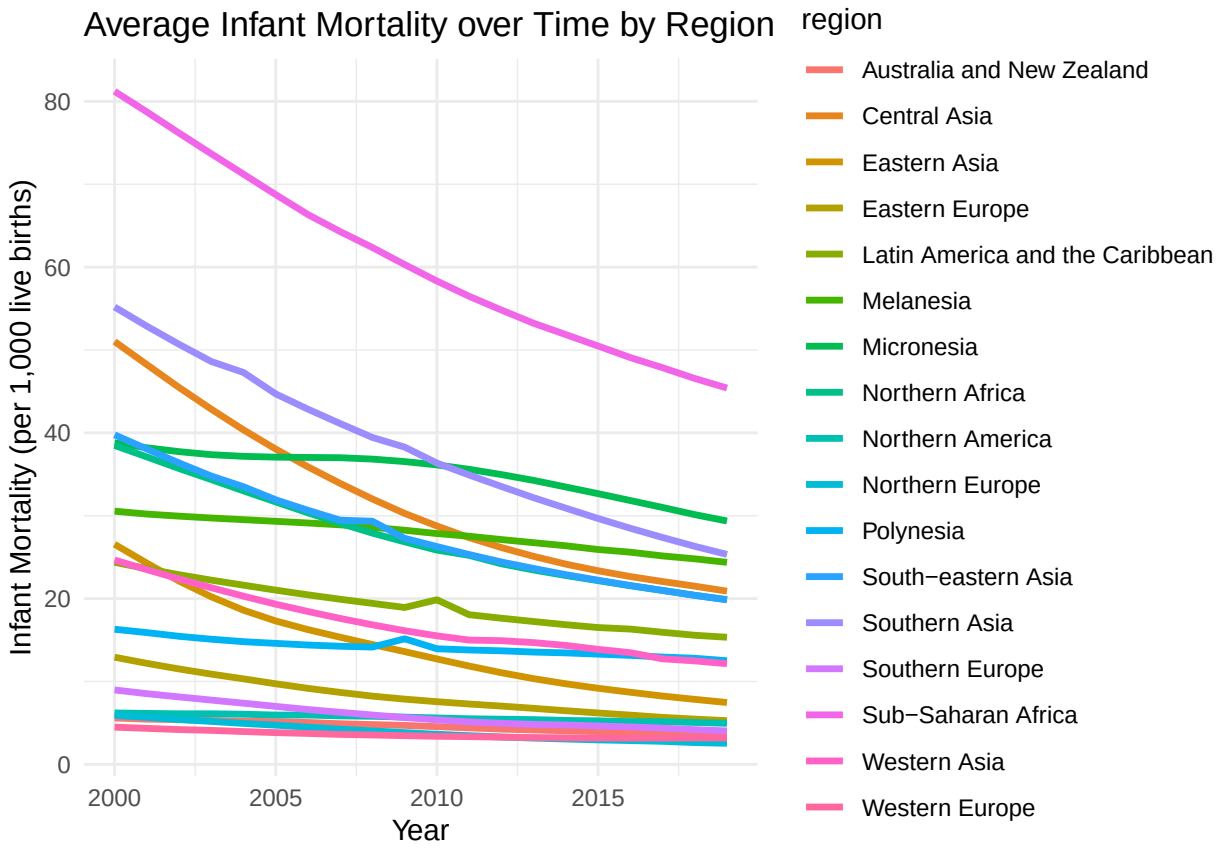
```
## Warning: Removed 426 rows containing non-finite values (`stat_summary()`).
```



Insight: - Maternal mortality has been declining in most regions, with significant improvements observed in certain areas like Sub-Saharan Africa and Southern Asia.

```
# Infant mortality trends by region
ggplot(data, aes(x = year, y = infmor, color = region)) +
  geom_line(stat = "summary", fun = "mean", size = 1.2) +
  labs(title = "Average Infant Mortality over Time by Region", x = "Year", y = "Infant Mortality (per 1
  theme_minimal()
```

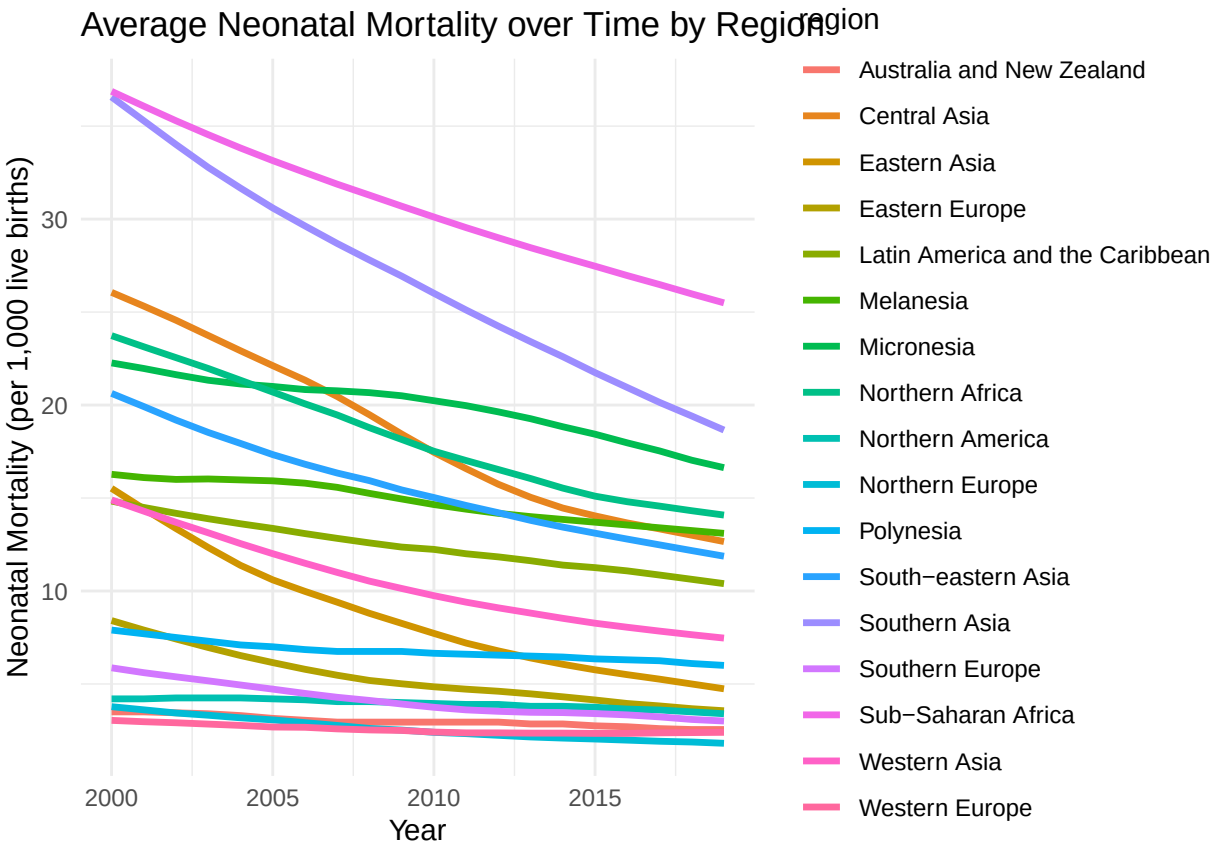
```
## Warning: Removed 20 rows containing non-finite values (`stat_summary()`).
```



Insight: - Similar to maternal mortality, infant mortality has been decreasing steadily across regions, although some regions still have higher rates than others.

```
# Neonatal mortality trends by region
ggplot(data, aes(x = year, y = neomor, color = region)) +
  geom_line(stat = "summary", fun = "mean", size = 1.2) +
  labs(title = "Average Neonatal Mortality over Time by Region", x = "Year", y = "Neonatal Mortality (p
  theme_minimal()
```

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



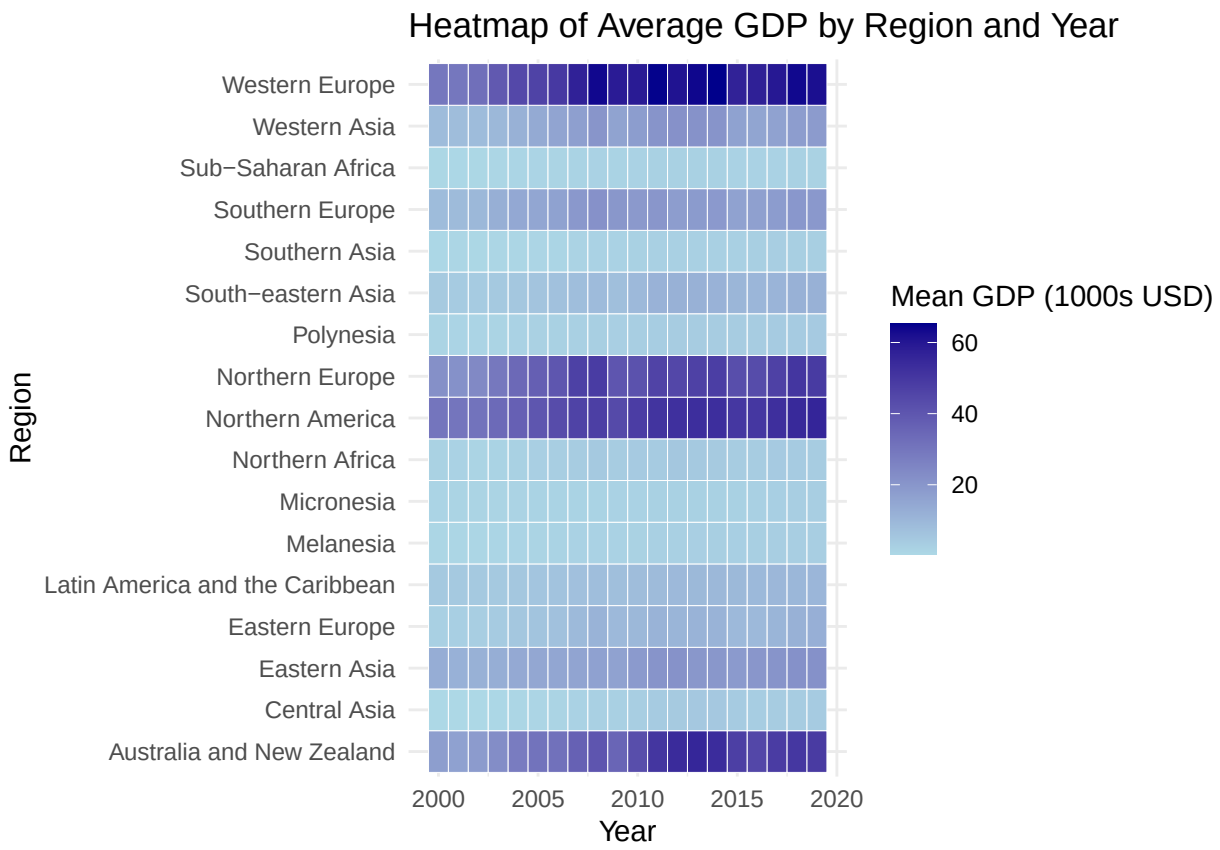
Insight: - Neonatal mortality is also declining but at a slower pace compared to maternal and infant mortality, indicating potential areas for further intervention in healthcare systems.

Heatmaps

```
# Heatmap for GDP by region and year
data_gdp_heatmap <- data %>%
  group_by(region, year) %>%
  summarize(mean_gdp = mean(gdp1000, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```

```
ggplot(data_gdp_heatmap, aes(x = year, y = region, fill = mean_gdp)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkblue", na.value = "grey50") +
  labs(title = "Heatmap of Average GDP by Region and Year", x = "Year", y = "Region", fill = "Mean GDP
  theme_minimal()
```

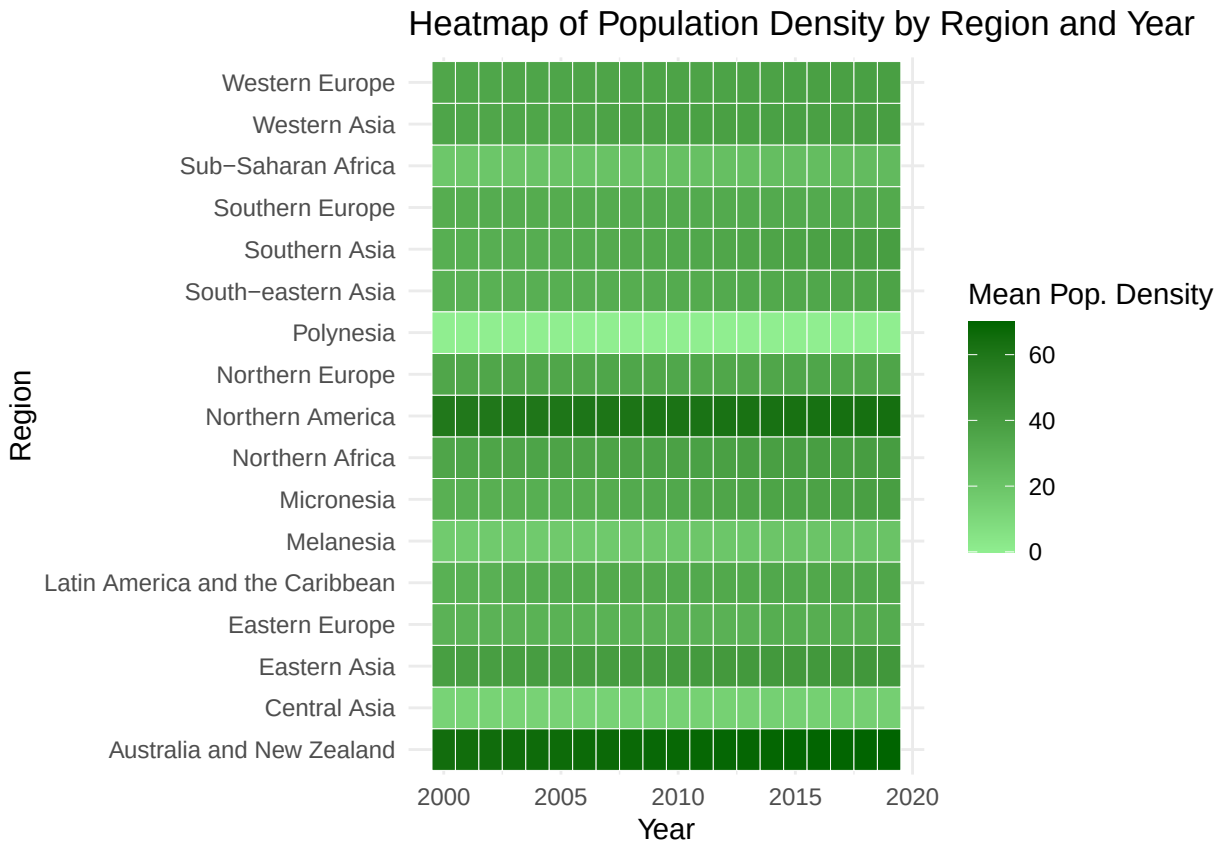


Insight: - The heatmap shows consistent GDP growth in regions like North America and Europe, while other regions like Africa and Southern Asia lag behind.

```
# Heatmap for Population Density by region and year
data_popdens_heatmap <- data %>%
  group_by(region, year) %>%
  summarize(mean_popdens = mean(popdens, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```

```
ggplot(data_popdens_heatmap, aes(x = year, y = region, fill = mean_popdens)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightgreen", high = "darkgreen", na.value = "grey50") +
  labs(title = "Heatmap of Population Density by Region and Year", x = "Year", y = "Region", fill = "Mea
  theme_minimal()
```

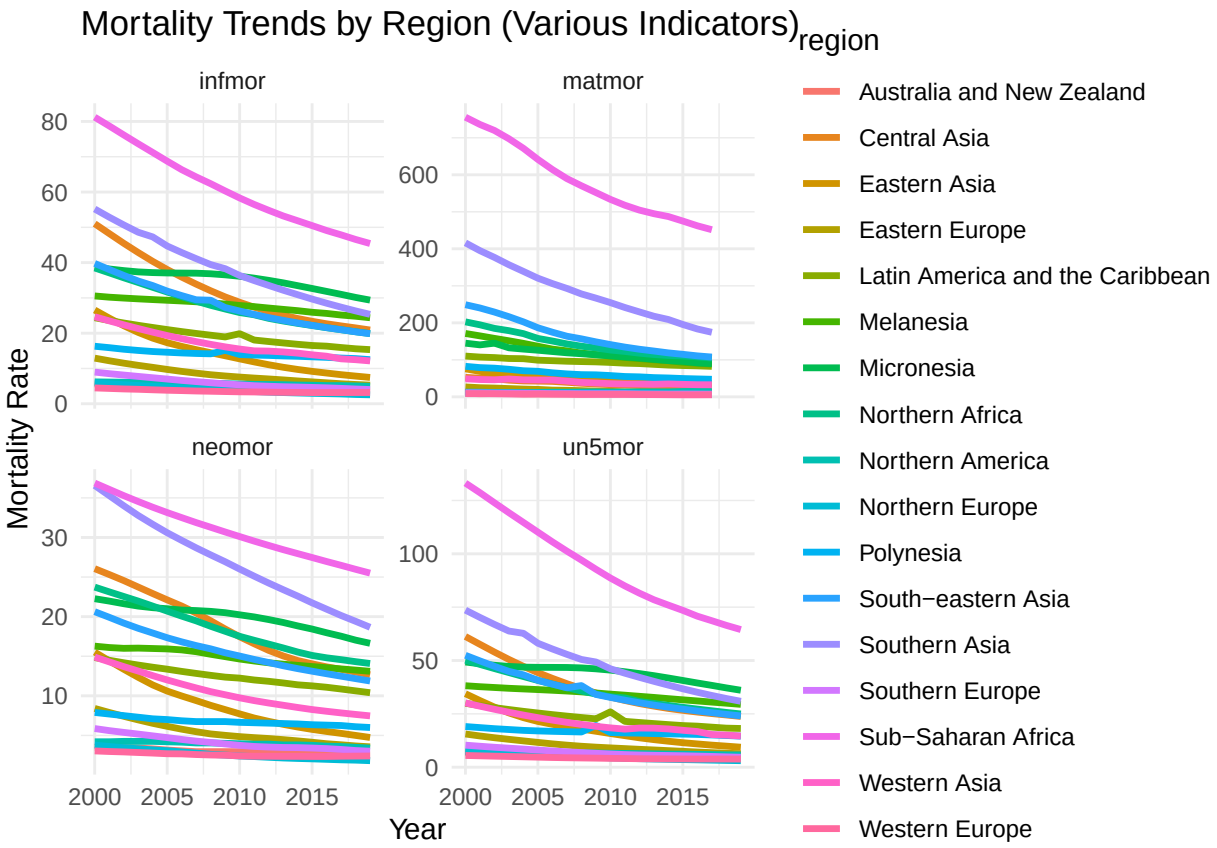


Insight: - The population density heatmap highlights regions with higher densities, such as Southern Asia and Sub-Saharan Africa, while other regions have more sparse populations.

```
# Comparing multiple mortality indicators by region
mortality_data <- data %>%
  select(region, year, matmor, infmor, neomor, un5mor) %>%
  gather(key = "mortality_type", value = "mortality_rate", matmor, infmor, neomor, un5mor)

ggplot(mortality_data, aes(x = year, y = mortality_rate, color = region)) +
  geom_line(stat = "summary", fun = "mean", size = 1.2) +
  facet_wrap(~mortality_type, scales = "free_y") +
  labs(title = "Mortality Trends by Region (Various Indicators)", x = "Year", y = "Mortality Rate") +
  theme_minimal()
```

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



Insight: - This visualization clearly shows the downward trend in mortality