EDA

John

Libraries and setup

```
here() starts at /Users/john/Desktop/~/ac


Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

-- Attaching packages --------------------------------------- tidyverse 1.3.1 --

v tibble  3.1.8     v purrr   1.0.1
v tidyr   1.3.0     v stringr 1.5.0
v readr   2.1.1     v forcats 0.5.1

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()

Warning: Some values were not matched unambiguously: Africa Eastern and Southern, Africa West

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

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

```
names(finaldata)
```

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

Missing Data Analysis

Count and proportion of NAs for each variable

```
var <- c("country_name", "ISO", "region", "year", "gdp1000", "OECD", "OECD2023",
               "popdens", "urban", "agedep", "male_edu", "temp", "rainfall1000", "matmor",
               "infmor", "neomor", "un5mor", "earthquake", "drought", "totdeath", "armconf1")

# Calculate the number and proportion of missing values for each variable
na_table <- data.frame(
  Variable = var,
  NA_Count = sapply(var, function(var) sum(is.na(finaldata[[var]]))),
  Proportion_NA = sapply(var, function(var) sum(is.na(finaldata[[var]])) / nrow(finaldata))
)

print(na_table)
```

```
                Variable NA_Count Proportion_NA
country_name country_name        0   0.000000000
ISO                  ISO        0   0.000000000
region            region        0   0.000000000
year                year        0   0.000000000
gdp1000          gdp1000       62   0.016666667
OECD                OECD        0   0.000000000
OECD2023        OECD2023        0   0.000000000
popdens          popdens       20   0.005376344
urban              urban       20   0.005376344
agedep            agedep        0   0.000000000
```

```
male_edu         male_edu       20   0.005376344
temp                 temp       20   0.005376344
rainfall1000 rainfall1000       20   0.005376344
matmor             matmor      426   0.114516129
infmor             infmor       20   0.005376344
neomor             neomor       20   0.005376344
un5mor             un5mor       20   0.005376344
earthquake     earthquake        0   0.000000000
drought           drought        0   0.000000000
totdeath         totdeath         0   0.000000000
armconf1         armconf1         0   0.000000000
```

Not sure which missing mechanism each variable is –MCAR, MAR, MNAR

```
finaldata |>
  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
 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         matmor           infmor           neomor
```

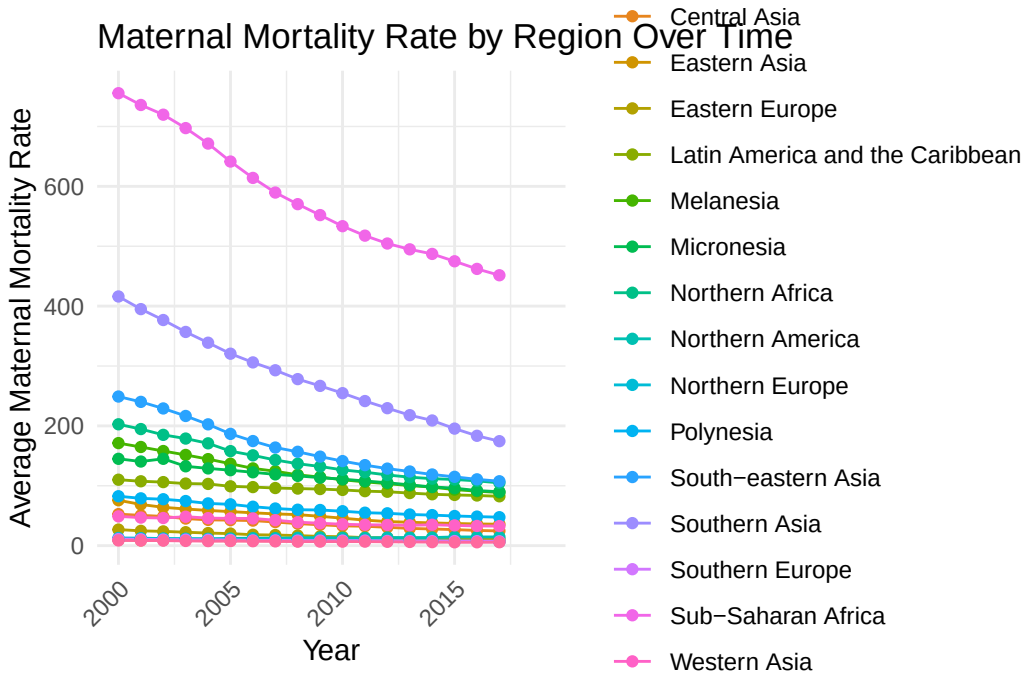
```
 Min.   :0.01993   Min.   :   2.0   Min.   :  1.60   Min.   : 0.80
 1st Qu.:0.59146   1st Qu.:  17.0   1st Qu.:  7.60   1st Qu.: 4.90
 Median :1.01288   Median :  66.0   Median : 18.90   Median :12.10
 Mean   :1.20216   Mean   : 210.6   Mean   : 28.90   Mean   :16.18
 3rd Qu.:1.68706   3rd Qu.: 299.8   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
     un5mor          earthquake         drought          totdeath
 Min.   :  2.00   Min.   :0.00000   Min.   :0.00000   Min.   :    0.0
 1st Qu.:  9.00   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:    0.0
 Median : 22.20   Median :0.00000   Median :0.00000   Median :    0.0
 Mean   : 40.50   Mean   :0.08737   Mean   :0.08333   Mean   :  361.1
 3rd Qu.: 61.33   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:    2.0
 Max.   :224.90   Max.   :1.00000   Max.   :1.00000   Max.   :78644.0
 NA's   :20
    armconf1
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.1892
 3rd Qu.:0.0000
 Max.   :1.0000
```

Plotting matmor, infmor, neomor, un5mor,

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

```
Warning: Removed 34 rows containing missing values (`geom_line()`).
```

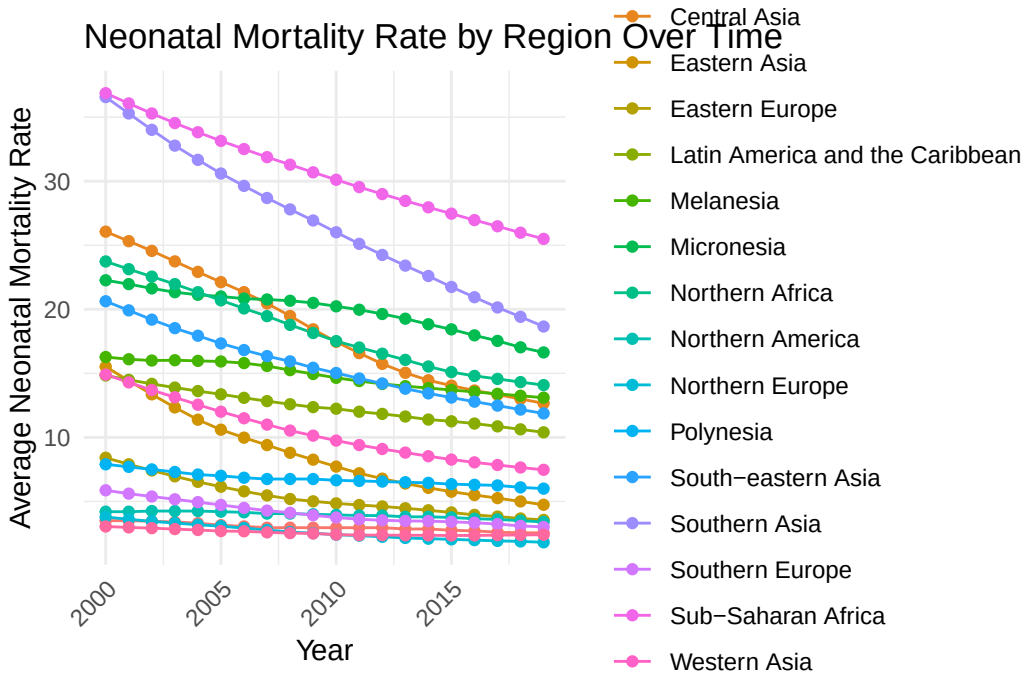
```
Warning: Removed 34 rows containing missing values (`geom_point()`).
```




```
# Group by region and year, and summarize the neomor
grouped_data <- finaldata %>%
  group_by(region, year) %>%
  summarise(neomor = mean(neomor, na.rm = TRUE))  # Summarize neomor by region and year
```

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

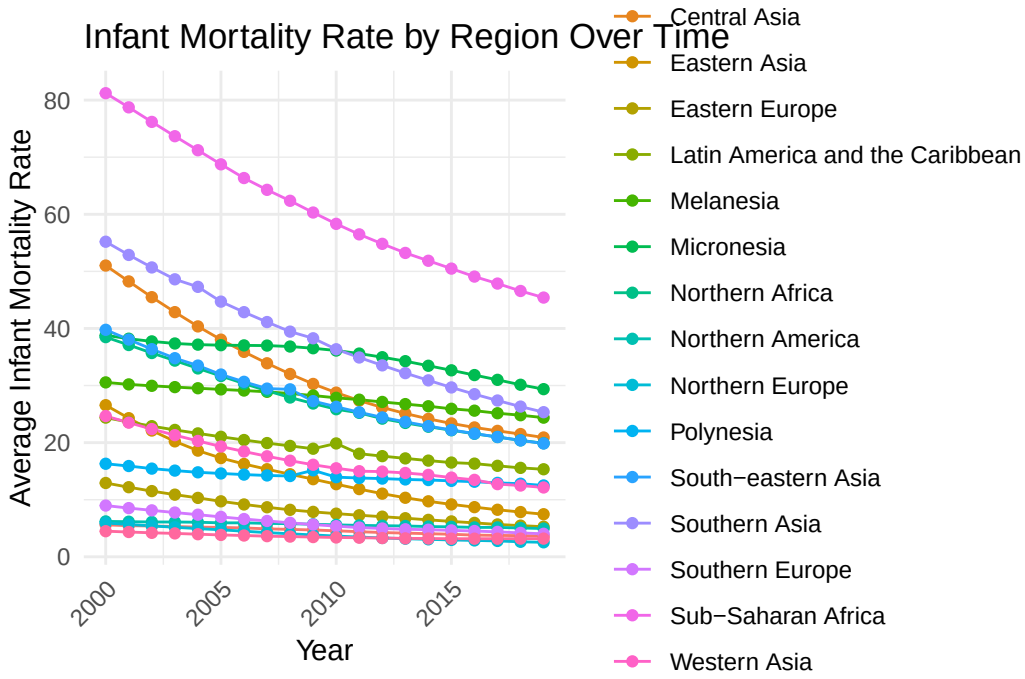
```
# Create the plot
ggplot(grouped_data, aes(x = year, y = neomor, group = region, color = region)) +
  geom_line() +
  geom_point() +
  labs(title = "Neonatal Mortality Rate by Region Over Time",
       x = "Year",
       y = "Average Neonatal Mortality Rate") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Tilting x-axis labels for reada
```




```
# Group by region and year, and summarize the infmor
grouped_data <- finaldata %>%
  group_by(region, year) %>%
  summarise(infmor = mean(infmor, na.rm = TRUE))  # Summarize infmor by region and year
```

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

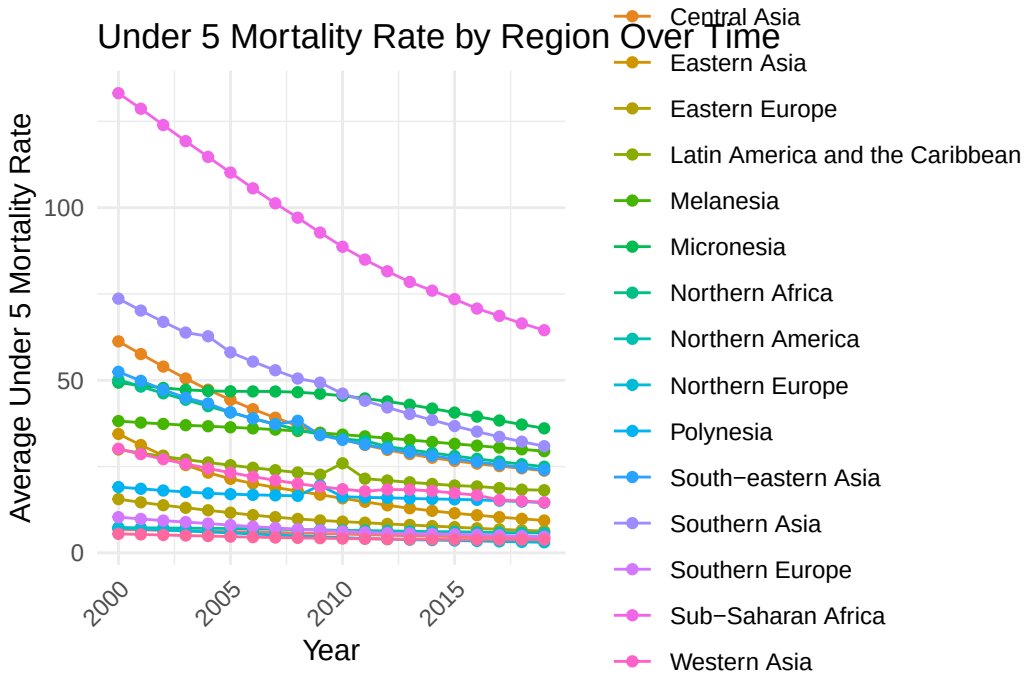
```
# Create the plot
ggplot(grouped_data, aes(x = year, y = infmor, group = region, color = region)) +
  geom_line() +
  geom_point() +
  labs(title = "Infant Mortality Rate by Region Over Time",
       x = "Year",
       y = "Average Infant Mortality Rate") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Tilting x-axis labels for reada
```




```
# Group by region and year, and summarize the un5mor
grouped_data <- finaldata %>%
  group_by(region, year) %>%
  summarise(un5mor = mean(un5mor, na.rm = TRUE))  # Summarize un5mor by region and year
```

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

```
# Create the plot
ggplot(grouped_data, aes(x = year, y = un5mor, group = region, color = region)) +
  geom_line() +
  geom_point() +
  labs(title = "Under 5 Mortality Rate by Region Over Time",
       x = "Year",
       y = "Average Under 5 Mortality Rate") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Tilting x-axis labels for reada
```



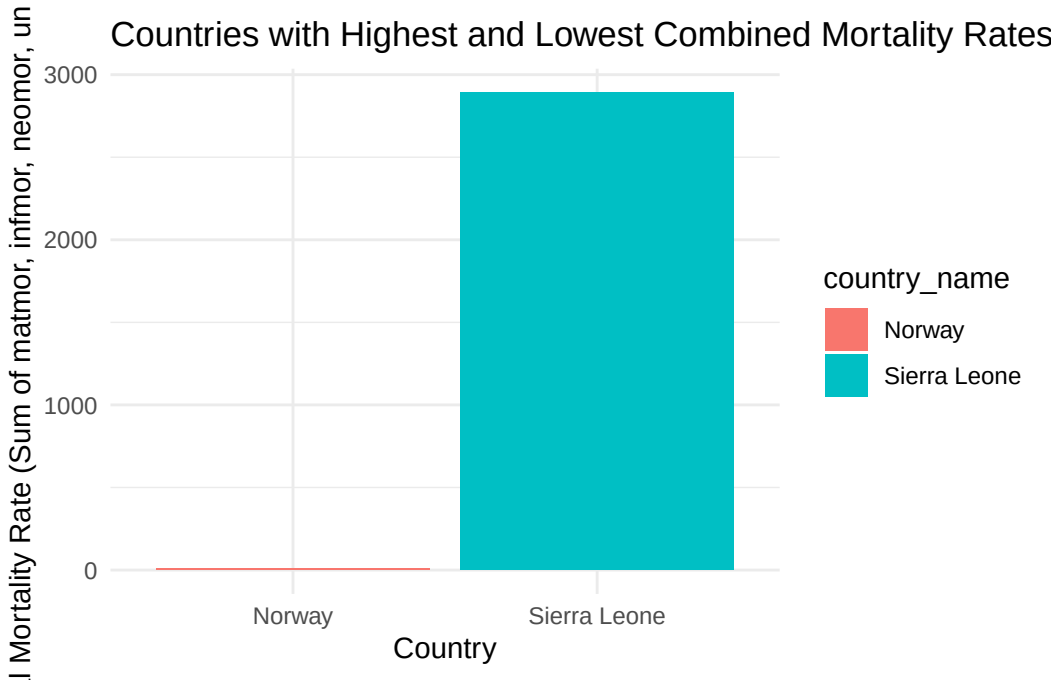
In general, all four mortality rates seems to be going down across all regions by year.

```
# Summing the four mortality rates for each country
finaldata <- finaldata %>%
  mutate(total_mortality = matmor + infmor + neomor + un5mor)

# Find the countries with the highest and lowest total mortality rate
max_mortality <- finaldata %>% filter(total_mortality == max(total_mortality, na.rm = TRUE))
min_mortality <- finaldata %>% filter(total_mortality == min(total_mortality, na.rm = TRUE))

# Combine the two into a single data frame
extreme_mortality <- bind_rows(max_mortality, min_mortality)

# Plotting the result
ggplot(extreme_mortality, aes(x = country_name, y = total_mortality, fill = country_name)) +
  geom_bar(stat = "identity") +
  labs(title = "Countries with Highest and Lowest Combined Mortality Rates",
       x = "Country",
       y = "Total Mortality Rate (Sum of matmor, infmor, neomor, un5mor)") +
  theme_minimal()
```



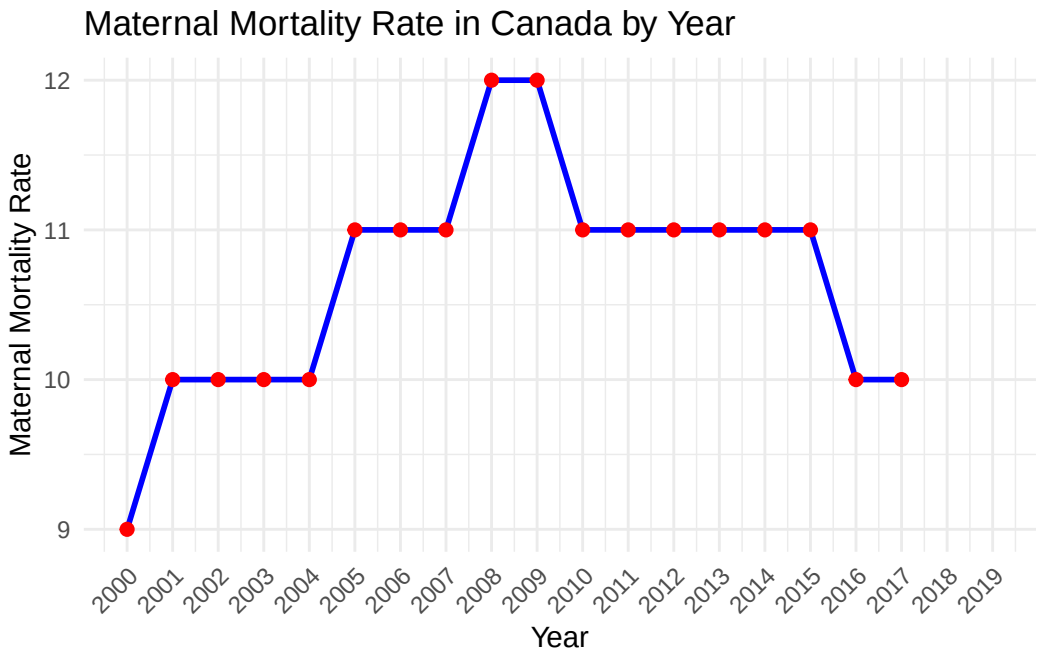

```
canada_data <- finaldata %>%
  filter(country_name == "Canada") %>%
  select(year, matmor)

ggplot(canada_data, aes(x = year, y = matmor)) +
  geom_line(color = "blue", size = 1) +  # Line plot for the trend
  geom_point(color = "red", size = 2) +  # Adding points to the line
  labs(title = "Maternal Mortality Rate in Canada by Year",
       x = "Year",
       y = "Maternal Mortality Rate") +
  theme_minimal() +
  scale_x_continuous(breaks = seq(min(canada_data$year), max(canada_data$year), by = 1)) + #
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
Warning: Removed 2 rows containing missing values (`geom_line()`).
```

```
Warning: Removed 2 rows containing missing values (`geom_point()`).
```


Maternal Mortality Rate in Canada by Year
Maternal Mortality Rate
12
11
10
9
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Year