EDA-Gender

PSTAT 296A

2025-10-17

Import packages

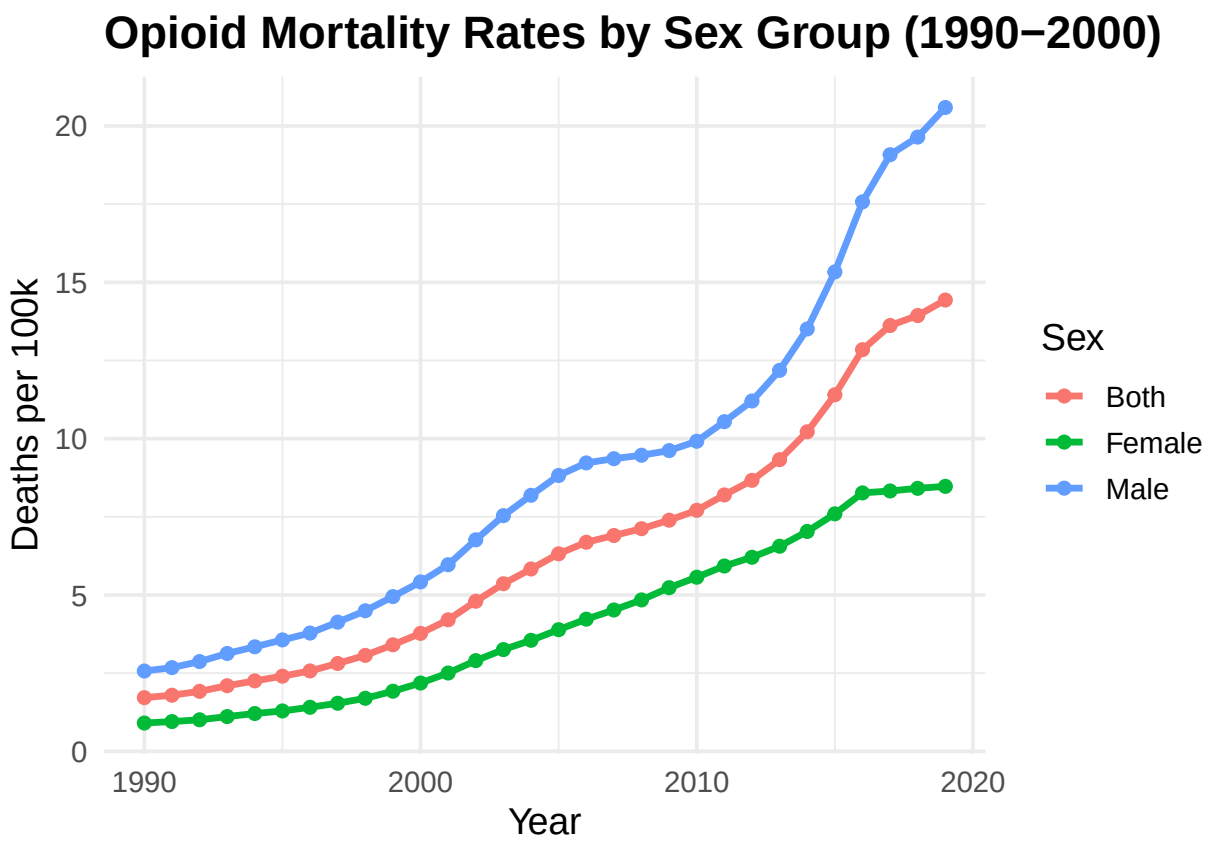
```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tseries)
library(forecast)
library(ggfortify)
library(strucchange)
```

EDA - Gender

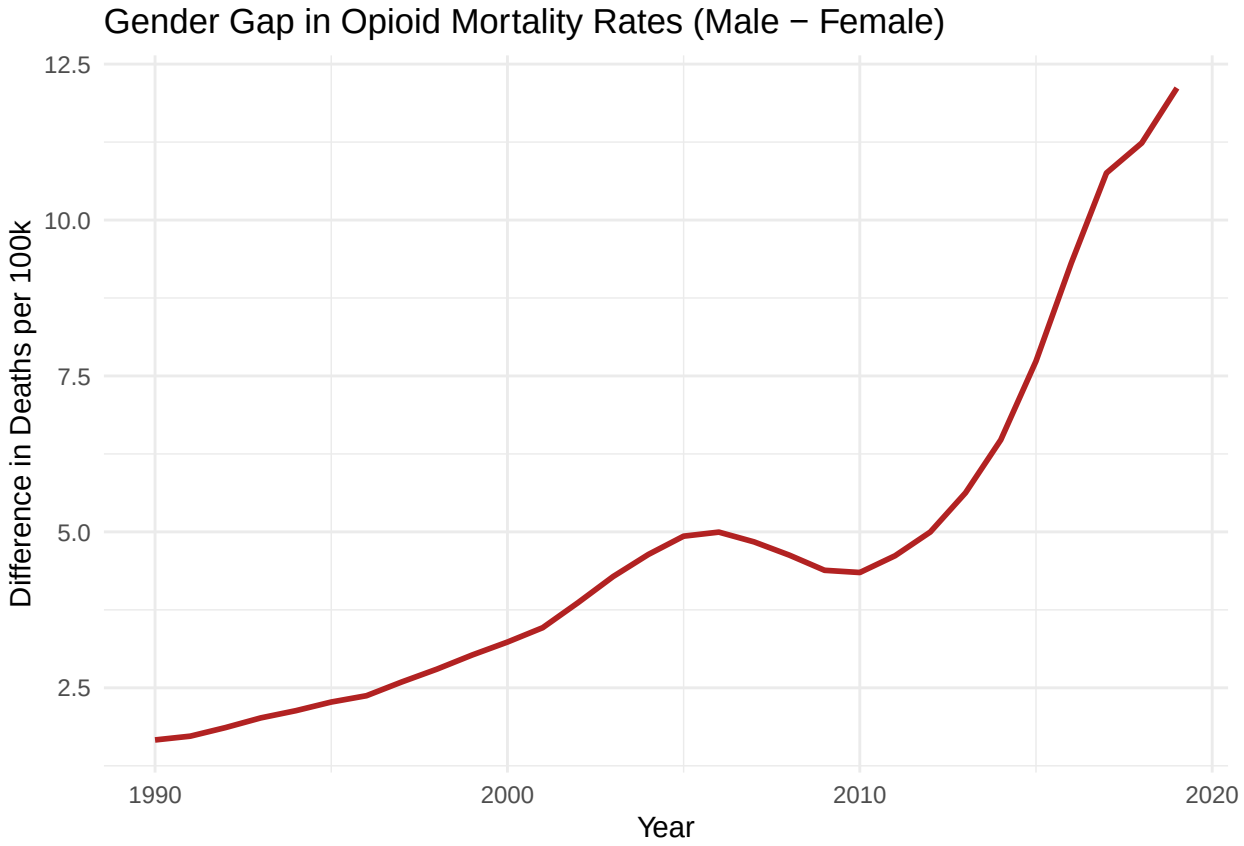
```
# Analyze age distribution (US opioid rate age.csv)
sex_data <- read.csv("Data/US opioid rate gender.csv")

# Check unique values of age bins
unique(sex_data$sex)
```

```
## [1] "Male"   "Female" "Both"
```

```
# Plot the distribution of opioid death rates by age
ggplot(sex_data, aes(x = year, y = val, color = sex)) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 1.8) +
  labs(
    title = "Opioid Mortality Rates by Sex Group (1990-2000)",
    x = "Year",
    y = "Deaths per 100k",
    color = "Sex"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "right",
    plot.title = element_text(face = "bold")
  )
```




```r
# Plot the gap between male and female mortality rate
sex_gap <- sex_data %>%
  filter(sex %in% c("Male","Female")) %>%
  group_by(year, sex) %>%
  pivot_wider(
    id_cols = year,
    names_from = sex,
    values_from = val
  ) %>%
  mutate(
    gap = Male - Female
  )

ggplot(sex_gap, aes(x = year, y = gap)) +
  geom_line(color = "firebrick", linewidth = 1) +
  labs(
    title = "Gender Gap in Opioid Mortality Rates (Male - Female)",
    x = "Year",
    y = "Difference in Deaths per 100k"
  ) +
  theme_minimal()
```

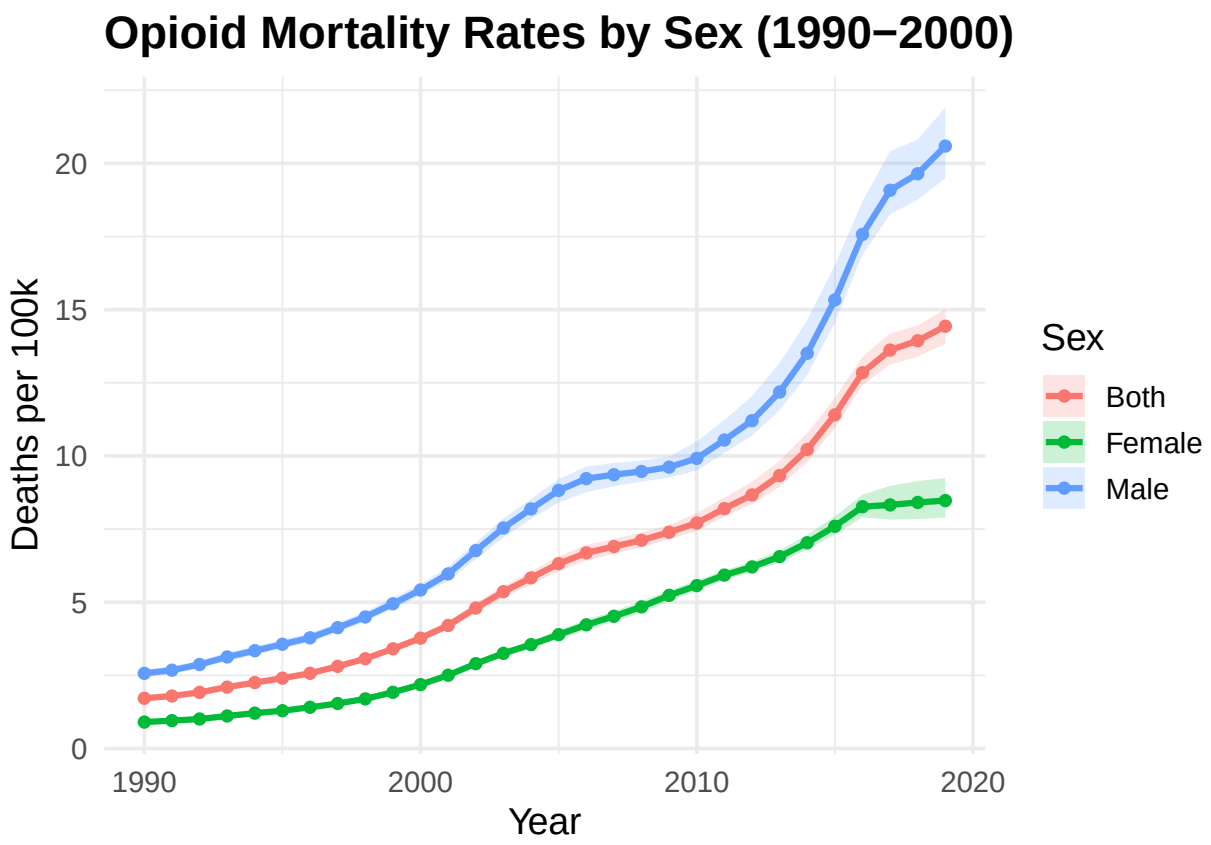


Between 1990 and 2000, mortality rate increase gradually.After 2000, the rates accelerated more steeply, especially after 2010.

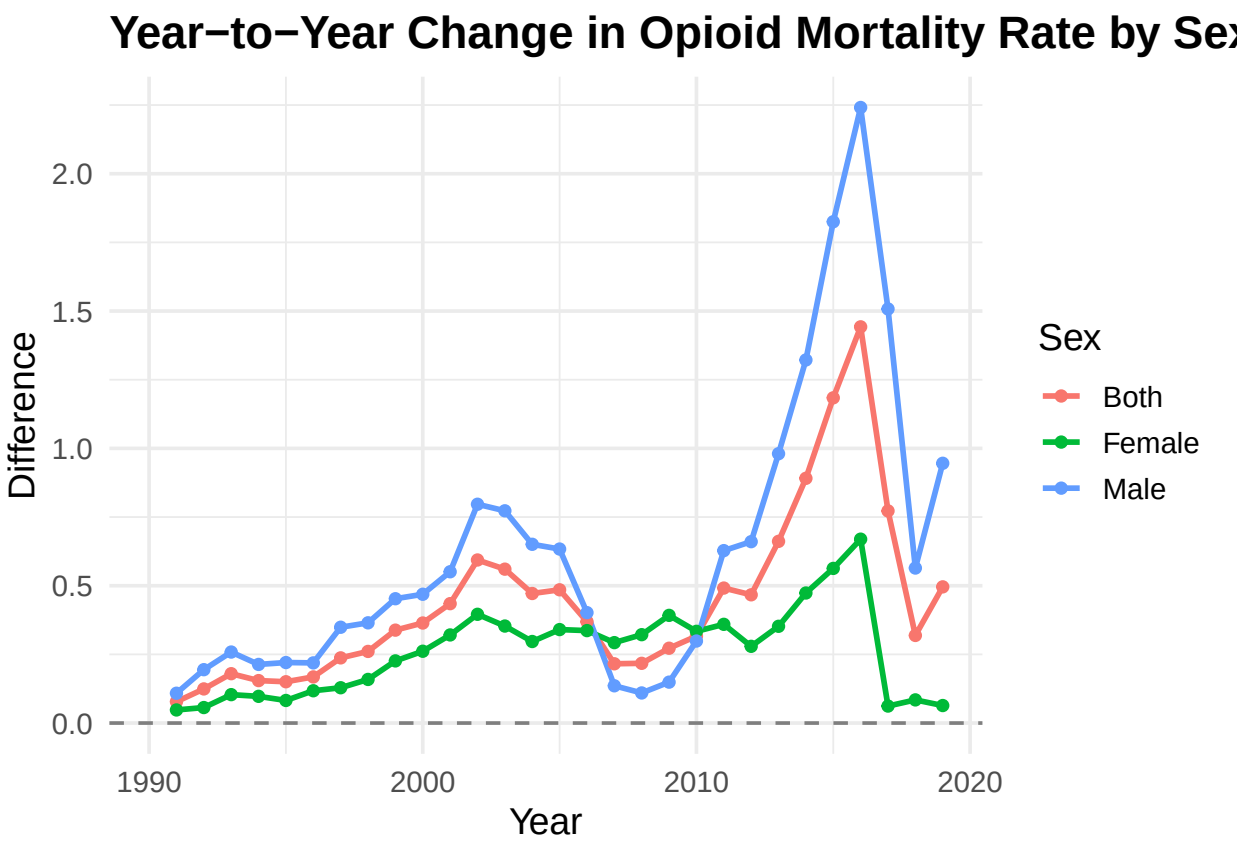
Males consistenly experience higher mortality rate that females through out the entire period. The difference between two group widens over time, especially after 2010. Male mortality rate rise at different pace overtime.

Females display a steadier and more linear increase from 1990 to around 2016. After 2016, the female mortality curve flattens, suggesting a stabilization in opioid-related deaths among women, while male rates continue to rise sharply.

```
# Add highlights around the confidence interval
ggplot(sex_data, aes(x = year, y = val, color = sex, fill = sex)) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.2, color = NA) +
  geom_line(linewidth = 1.1) +
  geom_point(size = 1.5) +
  labs(
    title = "Opioid Mortality Rates by Sex (1990-2000)",
    x = "Year",
    y = "Deaths per 100k",
    color = "Sex",
    fill = "Sex"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "right",
    plot.title = element_text(face = "bold")
  )
```



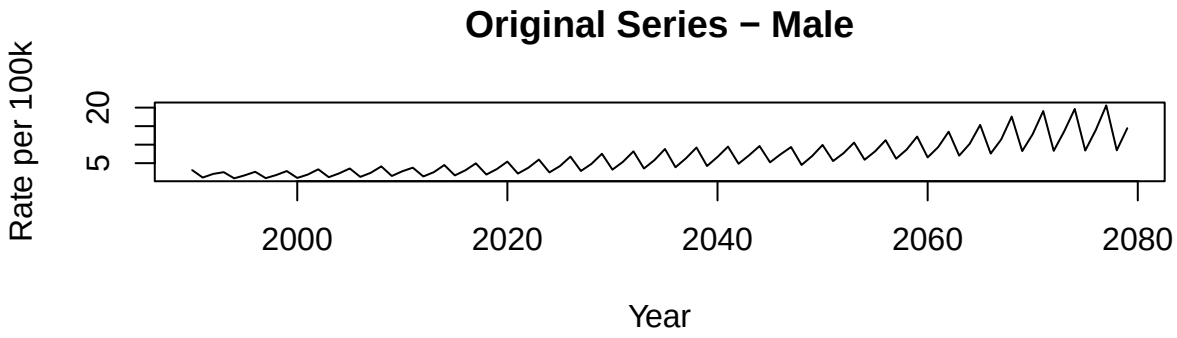

```r
# Compute year-to-year percent change by sex
sex_diff <- sex_data %>%
  group_by(sex) %>%
  arrange(year, .by_group = TRUE) %>%
  mutate(
    pct_change = (val - lag(val))  # Remvoe 100*
  ) %>%
  ungroup()

ggplot(sex_diff, aes(x = year, y = pct_change, color = sex)) +
  geom_line(linewidth = 1) +
  geom_point(size = 1.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray50") +
  labs(
    title = "Year-to-Year Change in Opioid Mortality Rate by Sex",
    x = "Year",
    y = "Difference",
    color = "Sex"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    legend.position = "right",
    plot.title = element_text(face = "bold")
  )
```

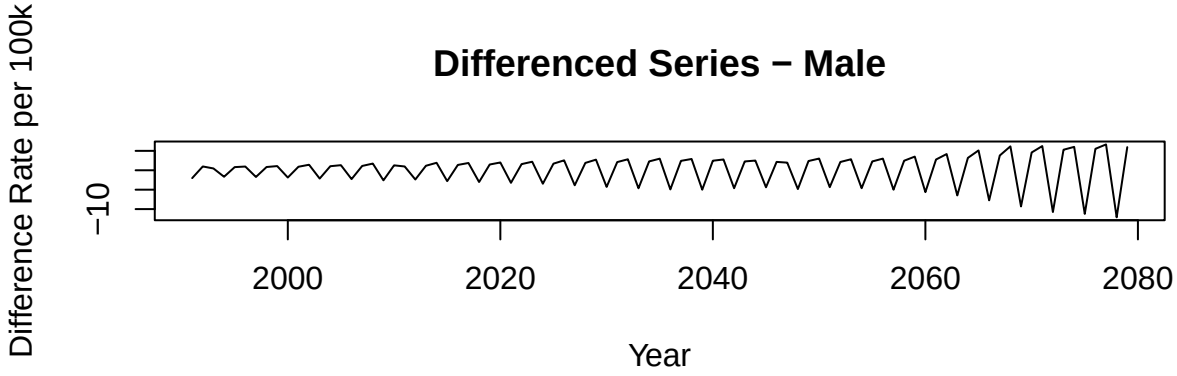



```r
# Loop through each age group
for (group in unique(sex_data$sex)) {

  cat("\n----------------------------------------\n")
  cat("Results for:", group, "\n")

  # Create time series
  ts_obj <- ts(sex_data$val, start = min(sex_data$year), frequency = 1)

  # Run ADF test on original
  adf_orig <- adf.test(ts_obj)
  cat("ADF (original series) p-value:", round(adf_orig$p.value, 4), "\n")

  # Difference the series
  ts_diff <- diff(ts_obj)

  # Run ADF test on differenced series
  adf_diff <- adf.test(ts_diff)
  cat("ADF (differenced series) p-value:", round(adf_diff$p.value, 4), "\n")

  # Optional: Plot both series
  par(mfrow = c(2, 1))  # two plots per group
  plot(ts_obj, main = paste("Original Series -", group),
       ylab = "Rate per 100k", xlab = "Year")
  plot(ts_diff, main = paste("Differenced Series -", group),
       ylab = "Difference Rate per 100k", xlab = "Year")
```

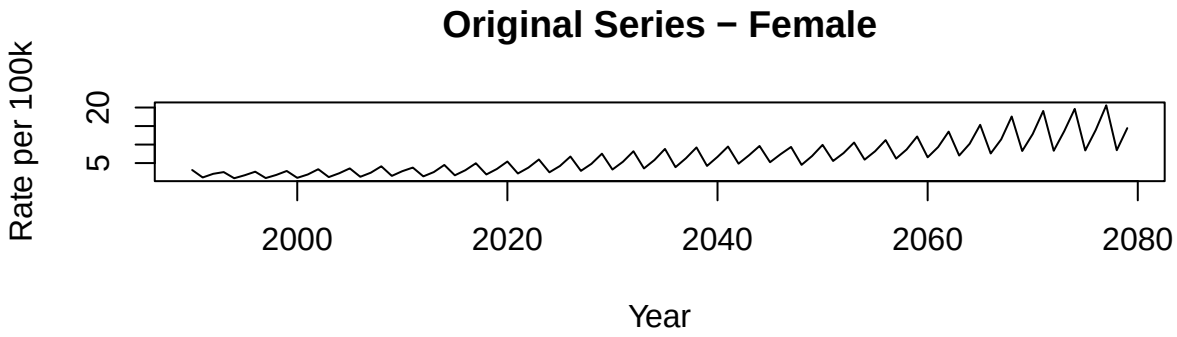
```
  cat("---------------------------------------\n")
}
```

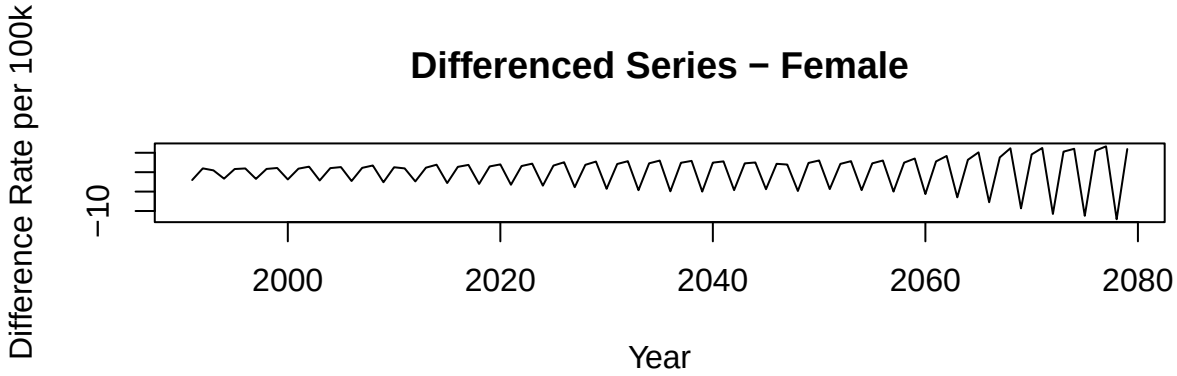
```
##
## -------------------------------------
## Results for: Male
## ADF (original series) p-value: 0.8739
## ADF (differenced series) p-value: 0.044
```



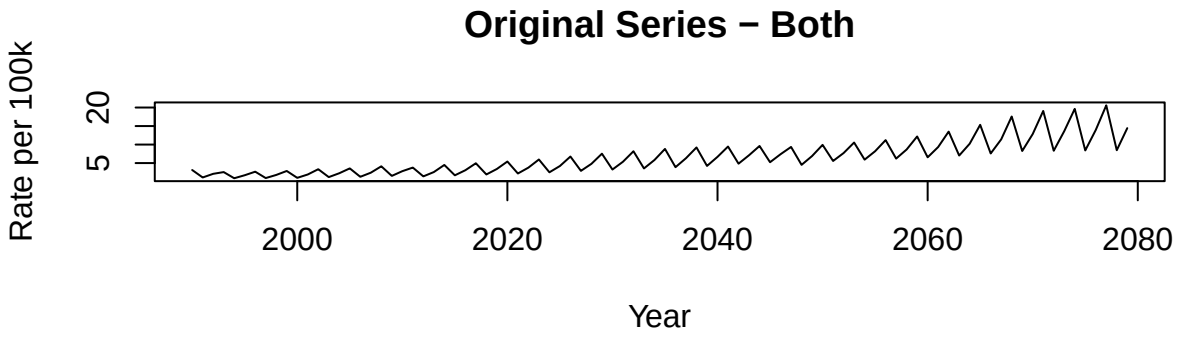


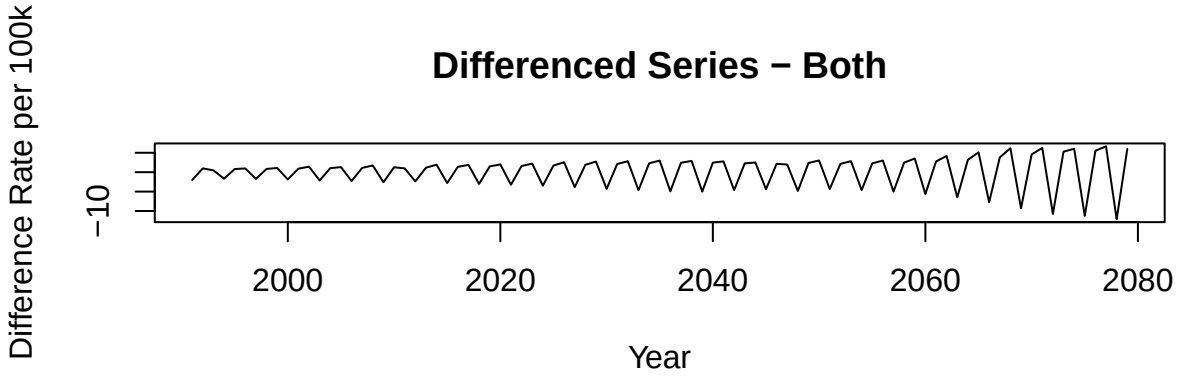

```
## -------------------------------------
##
## -------------------------------------
## Results for: Female
## ADF (original series) p-value: 0.8739
## ADF (differenced series) p-value: 0.044
```





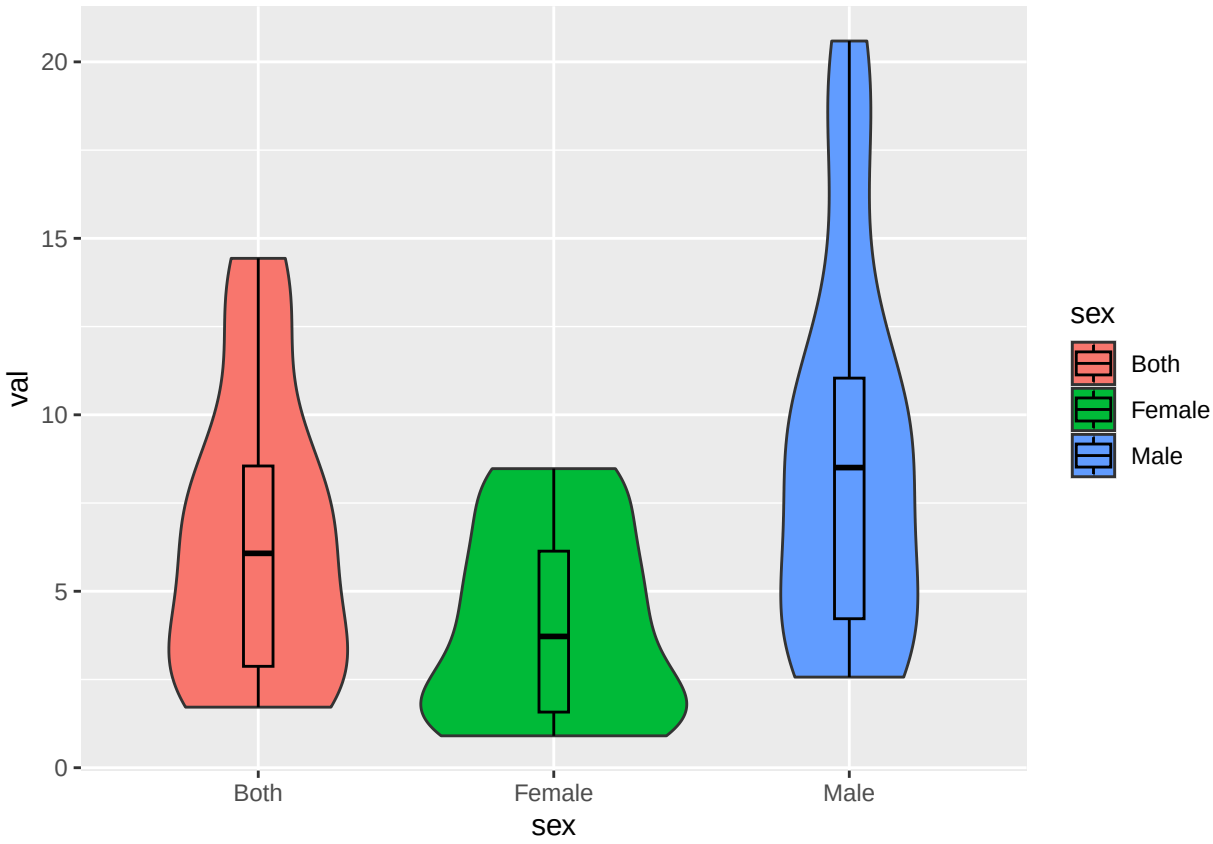

```
## ------------------------------------
##
## ------------------------------------
## Results for: Both
## ADF (original series) p-value: 0.8739
## ADF (differenced series) p-value: 0.044
```






```
## -------------------------------------
```

```r
ggplot(sex_data, aes(x = sex, y = val, fill = sex)) +
  geom_violin() +
  # geom_violin(fill = "lightblue") +
  geom_boxplot(width = 0.1, color = "black")   # optional overlay
```


20
15
10
5
0
val
Both
Female
Male
sex
sex
Both
Female
Male