Lab Exercise 1 App stat

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- 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 becor
```

```
dm <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1.txt", skip = 2, col_t
```

```
Warning: 494 parsing failures.
row    col               expected actual
108 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
109 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
110 Female no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
110 Male   no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
110 Total  no trailing characters      . 'https://www.prdh.umontreal.ca/BDLC/data/ont/Mx_1x1
... ...... ..................... ...... ................................................
See problems(...) for more details.
```

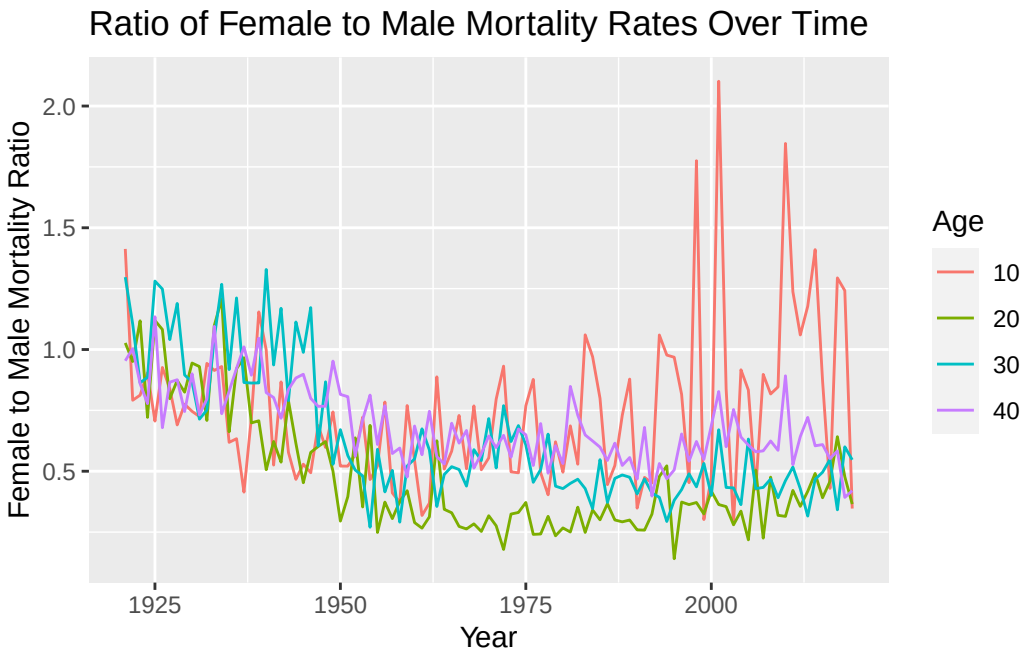
```
head(dm)
```

```
# A tibble: 6 x 5
   Year Age    Female    Male   Total
  <dbl> <chr>   <dbl>   <dbl>   <dbl>
1  1921 0     0.0978  0.129   0.114
2  1921 1     0.0129  0.0144  0.0137
3  1921 2     0.00521 0.00737 0.00631
4  1921 3     0.00471 0.00457 0.00464
5  1921 4     0.00461 0.00433 0.00447
6  1921 5     0.00372 0.00361 0.00367
```

1.

```r
# Calculate ratio of female to male mortality rates
mortality_ratio <- dm %>%
  filter(Age %in% c(10, 20, 30, 40)) %>%
  group_by(Year, Age) %>%
  mutate(ratio = Female / Male) %>%
  select(Year, Age, ratio)

# Plot
ggplot(mortality_ratio, aes(x = Year, y = ratio, color = as.factor(Age))) +
  geom_line() +
  labs(title = "Ratio of Female to Male Mortality Rates Over Time",
       x = "Year",
       y = "Female to Male Mortality Ratio",
       color = "Age") +
  # Change theme
  theme_gray()
```



2.

```
lowest_mortality_age <- dm %>%
  group_by(Year) %>%
  arrange(Female) %>%
  slice(1) %>%
  select(Year, Age, Female)
# Lowest Female Mortality Rate Each Year
lowest_mortality_age
```

```
# A tibble: 99 x 3
# Groups:   Year [99]
    Year Age     Female
   <dbl> <chr>    <dbl>
 1  1921 13    0.00176
 2  1922 104   0
 3  1923 105   0
 4  1924 14    0.00140
 5  1925 105   0
 6  1926 11    0.000942
 7  1927 9     0.00132
 8  1928 9     0.00105
```

```
 9  1929 10    0.00121
10  1930 13    0.00108
# i 89 more rows
```

3.

We can calculate the standard deviation of mortality rates by age by running this code.

```
std_dev_mortality <- dm %>%
  group_by(Age) %>%
  summarize(
    across(c(Female, Male, Total), sd,na.rm = TRUE)
  )
```

```
Warning: There was 1 warning in `summarize()`.
i In argument: `across(c(Female, Male, Total), sd, na.rm = TRUE)`.
i In group 1: `Age = "0"`.
Caused by warning:
! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
Supply arguments directly to `.fns` through an anonymous function instead.

  # Previously
  across(a:b, mean, na.rm = TRUE)

  # Now
  across(a:b, \(x) mean(x, na.rm = TRUE))
```

```
std_dev_mortality
```

```
# A tibble: 111 x 4
   Age     Female     Male    Total
   <chr>    <dbl>    <dbl>    <dbl>
 1 0     0.0256   0.0330   0.0294
 2 1     0.00352  0.00396  0.00374
 3 10    0.000474 0.000561 0.000509
 4 100   0.0928   0.138    0.0729
 5 101   0.125    0.158    0.0995
 6 102   0.143    0.214    0.114
 7 103   0.252    0.371    0.208
 8 104   0.449    1.01     0.363
```

```
 9 105    1.27     1.29     1.27
10 106    1.21     1.13     1.20
# i 101 more rows
```

4.

As we can see in the graph, male mortality rates consistently higher than female rates throughout the period observed.

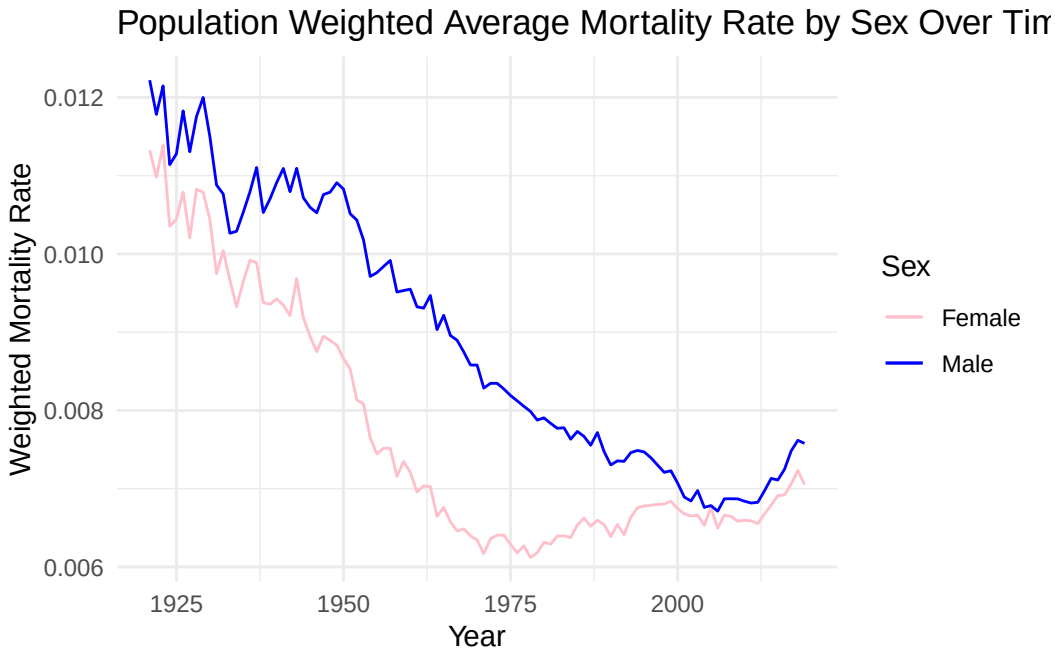
```
dm2 <- read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt", skip = 1)
```

```
-- Column specification --------------------------------------------------------
cols(
  Year = col_double(),
  Age = col_character(),
  Female = col_double(),
  Male = col_double(),
  Total = col_double()
)
```

```
# Calculate the population-weighted average mortality rate
weighted_mortality <- dm %>%
  left_join(dm2, by = c("Year", "Age")) %>%
  # Drop missing values
  drop_na() %>%
  group_by(Year) %>%
  summarize(
     weighted_average_mortality_female = weighted.mean(Female.x, w = Female.y, na.rm = TRU
     weighted_average_mortality_male = weighted.mean(Male.x, w = Male.y, na.rm = TRUE)
   )

# Plot the results
ggplot(weighted_mortality, aes(x = Year)) +
  geom_line(aes(y = weighted_average_mortality_female, color = "Female")) +
  geom_line(aes(y = weighted_average_mortality_male, color = "Male")) +
  scale_color_manual(values = c("Female" = "pink", "Male" = "blue")) +
  labs(title = "Population Weighted Average Mortality Rate by Sex Over Time",
       x = "Year",
       y = "Weighted Mortality Rate",
       color = "Sex") +
```

```
  theme_minimal()
```



5.

For a simple linear regression model with logged mortality rates as the outcome and age as the covariate, the notation of the simple linear regression is:

$$log(MortalityRate) = \beta_0 + \beta_1 Age + \epsilon$$

The output of the summary suggests that $\beta_0 = -10.062281$ and $\beta_1 = 0.086891$. The positive coefficient for Age suggests that the log of the mortality rate increases as age increases, which implies that the mortality rate itself also increases exponentially with age. Given the context of mortality data, this result is consistent with general expectations: as age increases, the risk of mortality typically increases.

```
# Run the linear regression with logged mortality rates

female_data <- dm %>%
  # Transform data type since Age is Character
  mutate(Age = as.integer(Age)) %>%
  # There is 110+, which can't be converted to integer. This is coverted to NA.
  # Since we only care about Age under 106, we remove this.
  drop_na() %>%
```

```
    filter(Age < 106, Year == 2000)
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `Age = as.integer(Age)`.
Caused by warning:
! NAs introduced by coercion
```

```
model <- lm(log(Female) ~ Age, data = female_data)
summary(model)
```

```
Call:
lm(formula = log(Female) ~ Age, data = female_data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9692 -0.3194 -0.1341  0.2734  4.7993

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.062281   0.121345  -82.92   <2e-16 ***
Age           0.086891   0.001997   43.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6291 on 104 degrees of freedom
Multiple R-squared:  0.9479,    Adjusted R-squared:  0.9474
F-statistic:  1893 on 1 and 104 DF,  p-value: < 2.2e-16
```