Week 1 lab: Introduction to Tidyverse

Hainan Xu

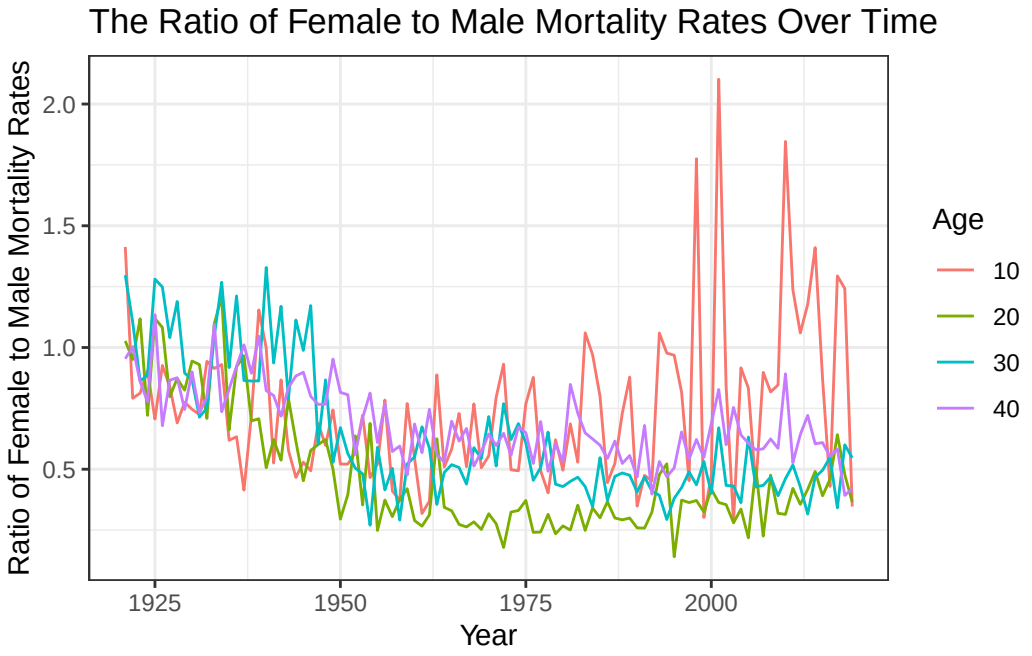
Lab Exercises

1. Plot the ratio of female to male mortality rates over time for ages 10,20,30 and 40 (different color for each age) and change the theme

```
library(tidyverse)
library(ggplot2)
```

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

```
dm1<-dm |> mutate(fm_ratio=Female/Male) |>
  filter(Age %in% c(10,20,30,40)) |>
  select(Year,Age,fm_ratio)
dm1 |> ggplot(aes(x=Year, y=fm_ratio, color=Age)) +
  geom_line() +
  labs(title="The Ratio of Female to Male Mortality Rates Over Time ",y="Ratio of Female t
  theme_bw()
```



2. Find the age that has the lowest female mortality rate each year

The dataframe below contains the ages that has the lowest female mortality rate each year. In some years, there are several ages that has the lowest female mortality rate.

```
dm2 <- dm |> select(-Male,-Total) |>
  group_by(Year) |>
  filter(Female==min(Female,na.rm = TRUE))
head(dm2)
```

```
# A tibble: 6 x 3
# Groups:   Year [4]
   Year Age    Female
  <dbl> <chr>   <dbl>
1  1921 13    0.00176
2  1922 104   0
3  1922 105   0
4  1923 105   0
5  1923 106   0
6  1924 14    0.00140
```

3. Use the 'summarize(across())' syntax to calculate the standard deviation of mortality rates by age for the Male, Female and Total populations.

The output below show the first 6 rows of the standard deviation of mortality rates by age for male, female and total populations.

```
dm3 <- dm |> group_by(Age)|>
  summarize(across(Female:Total,~sd(.x,na.rm = TRUE)))
head(dm3)
```

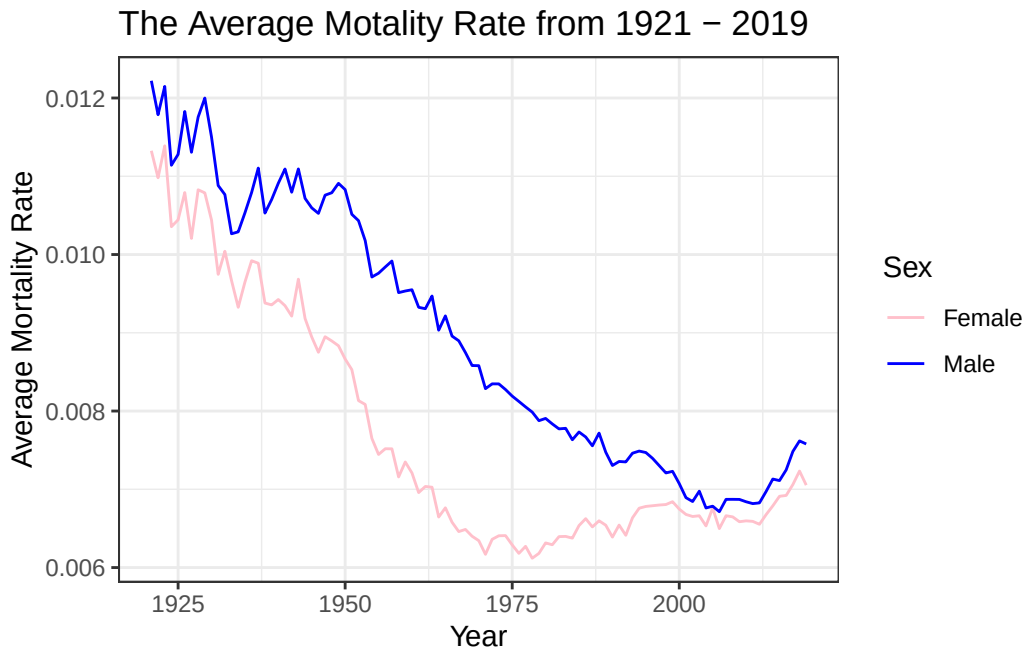
```
# A tibble: 6 x 4
  Age     Female     Male    Total
  <chr>    <dbl>    <dbl>    <dbl>
1 0      0.0256   0.0330   0.0294
2 1      0.00352  0.00396  0.00374
3 10     0.000474 0.000561 0.000509
4 100    0.0928   0.138    0.0729
5 101    0.125    0.158    0.0995
6 102    0.143    0.214    0.114
```

4. The Canadian HMD also provides population sizes over time. Use these to calculate the population weighted average mortality rate separately for males and females, for every year. Make a nice line plot showing the result (with meaningful labels/titles) and briefly comment on what you see (1 sentence). Hint: 'left_join' will probably be useful here.

Comment: The observed output indicates a general decreasing trend in the average mortality rates for both women and men over time, with a slight rebound since 2007; furthermore, males exhibit a consistently higher mortality rate compared to females.

```
#import the additional data
hmd=read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt",skip = 2, col_
#combine the data
dm4 <- dm |> left_join(hmd, by = c("Year","Age"))
dm4.1<- dm4 |> mutate(female_mortality=Female.x*Female.y,
                male_mortality=Male.x*Male.y) |>
  group_by(Year) |>
  summarise(#denominator of weighted average
            female_death_total=sum(female_mortality,na.rm = TRUE),
            male_death_total=sum(male_mortality, na.rm = TRUE),
            #numerator of weighted average
            female_total=sum(Female.y),
            male_total=sum(Male.y)
            ) |>
  mutate(Female_Average=female_death_total/female_total,
```

```
        Male_Average=male_death_total/male_total) |>
  select(Year,Female_Average, Male_Average)
#plot
dm4.1 <- dm4.1 |> pivot_longer(cols=c("Female_Average","Male_Average"), names_to = "Sex",
ggplot(dm4.1)+
  geom_line(aes(x=Year,y=Average_Mortality_Rate, color= Sex))+
  scale_color_manual(
    values = c("Female_Average" = "pink", "Male_Average" = "blue"),
    labels = c("Female", "Male")) +
  labs(title = "The Average Motality Rate from 1921 - 2019", y="Average Mortality Rate")+
    theme_bw()
```



5. Write down using appropriate notation, and run a simple linear regression with logged mortality rates as the outcome and age (as a continuous variable) as the covariate, using data for females aged less than 106 for the year 2000. Interpret the coefficient on age.

We fit the model

$$log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

where y_i is the mortality rate of the whole population in Ontario in 2000, x_i is the age, $\epsilon_i \sim (0, \sigma^2)$. According to the model summary shown below, the coefficient on age β_1 is 0.084,

which means on average, the mortality rate increases 8.4% for one year increase in age.

```
dm5<- dm |> mutate(Age=as.numeric(Age)) |>
  filter(Year ==2000, Age < 106)|>
  na.omit()
```

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

```
model<-lm(log(Total) ~ Age, data= dm5)
summary(model)
```

```
Call:
lm(formula = log(Total) ~ Age, data = dm5)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6641 -0.3662 -0.0879  0.2742  4.4168

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.644621   0.109453  -88.12   <2e-16 ***
Age          0.083984   0.001801   46.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5674 on 104 degrees of freedom
Multiple R-squared:  0.9543,    Adjusted R-squared:  0.9539
F-statistic:  2174 on 1 and 104 DF,  p-value: < 2.2e-16
```