STA2201-lab1

Kerry Hu

Lab Exercises

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

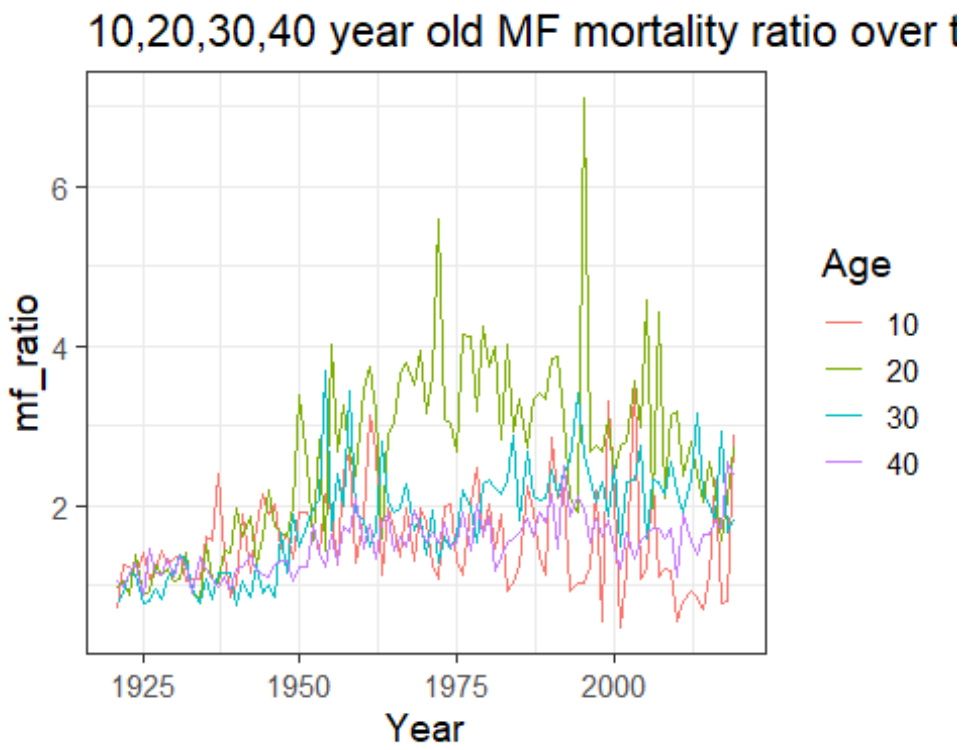
```
library(tidyverse)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2
──
✓ ggplot2 3.4.0      ✓ purrr   0.3.4
✓ tibble  3.1.8      ✓ dplyr   1.0.10
✓ tidyr   1.2.1      ✓ stringr 1.5.0
✓ readr   2.1.3      ✓ forcats 0.5.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

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

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

```
dp <- dm |>
      filter(Age==10|Age==20|Age==30|Age==40)|>
      mutate(mf_ratio = Male/Female)|>select(Year,Age,mf_ratio)
  dp|> ggplot(aes(x=Year,y=mf_ratio,color = Age))+geom_line()+labs(title =
"10,20,30,40 year old MF mortality ratio over time,
Ontario")+theme_bw(base_size = 14)
```



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

```
summary_max <- dm |> group_by(Year)|>select(Year,Age,Female)|>
summarize(max_mortality = max(Female, na.rm = TRUE))
df_list1 = list(dm,summary_max)
dmf<- df_list1 |> reduce(left_join, by='Year')
|>select(Year,Age,Female,max_mortality)
dmf <- dmf |>
    mutate_at(3, ~replace_na(.,0))
Year=rep(0,200)
Age=rep(0,200)
YearAge1<-data.frame(Year,Age)

j=1
for(i in 1:10989){
  if(dmf$Female[i]==dmf$max_mortality[i]) {YearAge1$Age[j]<-
dmf$Age[i];YearAge1$Year[j]<-dmf$Year[i];j=j+1}
  }
YearAge<-head(YearAge1,102)
YearAge

    Year  Age
1   1921  106
2   1922   98
3   1923  104
4   1924  107
5   1925   98
```

6	1926	106
7	1927	106
8	1928	104
9	1929	104
10	1930	105
11	1931	104
12	1932	105
13	1933	104
14	1934	106
15	1935	104
16	1936	106
17	1937	105
18	1938	104
19	1939	105
20	1940	104
21	1941	105
22	1942	104
23	1943	105
24	1944	98
25	1945	104
26	1946	105
27	1947	104
28	1948	99
29	1949	102
30	1950	102
31	1951	110+
32	1952	107
33	1953	106
34	1954	110+
35	1955	107
36	1956	110+
37	1957	107
38	1958	110+
39	1959	108
40	1960	107
41	1961	106
42	1962	108
43	1963	109
44	1964	109
45	1965	109
46	1966	105
47	1967	107
48	1968	97
49	1969	109
50	1970	107
51	1971	107
52	1972	107
53	1973	105
54	1974	107
55	1974	109

```
56  1975  108
57  1975 110+
58  1976  106
59  1976  108
60  1977  103
61  1978  109
62  1979  109
63  1980 110+
64  1981  107
65  1982  109
66  1983 110+
67  1984 110+
68  1985 110+
69  1986  109
70  1987  109
71  1988  108
72  1989  108
73  1990  108
74  1991  103
75  1992  108
76  1993  109
77  1994  109
78  1995  107
79  1996  109
80  1997  107
81  1998 110+
82  1999 110+
83  2000  106
84  2001 110+
85  2002  107
86  2003  109
87  2004  108
88  2005  108
89  2006 110+
90  2007  107
91  2008  109
92  2009 110+
93  2010  108
94  2011 110+
95  2012  109
96  2013 110+
97  2014 110+
98  2015 110+
99  2016 110+
100 2017 110+
101 2018 110+
102 2019 110+
```

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

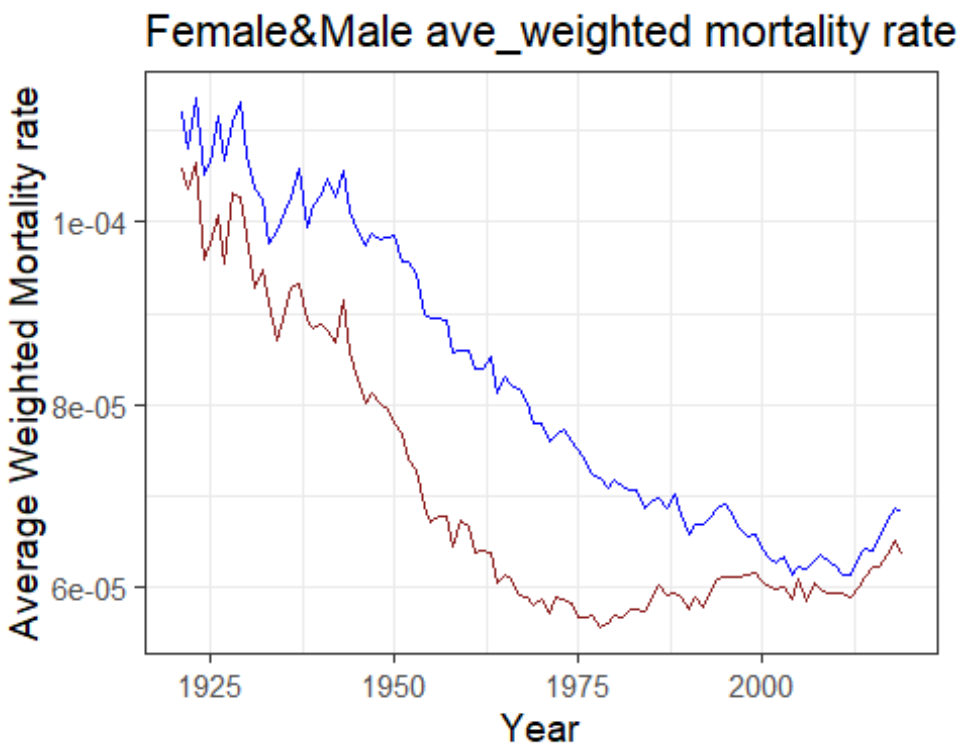
```
summary_SD <- dm |> group_by(Age)|>
  summarize(across(Female:Total, sd))
summary_SD

# A tibble: 111 × 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     NA        0.363
 9 105    NA        NA        NA
10 106    NA        NA        NA
# … with 101 more rows
```

4. The Canadian HMD also provides population sizes over time (https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt). 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.

```
df <-
read_table("https://www.prdh.umontreal.ca/BDLC/data/ont/Population.txt",skip
= 2, col_types = "dcddd")
df_list1 = list(dm,df)
dmf<- df_list1 |> reduce(left_join, by=c('Year','Age'))
dmf <-dmf |> mutate(weighed_F=Female.x*Female.y,weighed_M=Male.x*Male.y)|>
select(-Total.x,-Total.y)
dmfsum<- dmf |>group_by(Year) |>summarize(total_F=sum(Female.y, na.rm =
TRUE),total_M=sum(Male.y, na.rm = TRUE))
df_list2 = list(dmf,dmfsum)
dmf<- df_list2|> reduce(left_join, by='Year')
dmf

# A tibble: 10,989 × 10
    Year Age   Female.x  Male.x Female.y Male.y weighe…¹ weigh…² total_F
total_M
   <dbl> <chr>    <dbl>   <dbl>    <dbl>  <dbl>    <dbl>   <dbl>   <dbl>
<dbl>
 1  1921 0      0.0978  0.129     30157. 31530.   2948.   4070.   1.46e6
1.48e6
 2  1921 1      0.0129  0.0144    30391. 31319.    394.    452.   1.46e6
1.48e6
 3  1921 2      0.00521 0.00737   30962. 31785.    161.    234.   1.46e6
1.48e6
 4  1921 3      0.00471 0.00457   31306. 32031.    147.    146.   1.46e6
```

```
1.48e6
 5  1921 4      0.00461 0.00433   31364. 32046.    145.    139.   1.46e6
1.48e6
 6  1921 5      0.00372 0.00361   31175. 31847.    116.    115.   1.46e6
1.48e6
 7  1921 6      0.00265 0.00393   30808. 31466.     81.7   124.   1.46e6
1.48e6
 8  1921 7      0.00295 0.00351   30295. 30922      89.5   108.   1.46e6
1.48e6
 9  1921 8      0.00237 0.00285   29660. 30270.     70.4    86.4  1.46e6
1.48e6
10  1921 9      0.00198 0.00255   28923  29494.     57.4    75.3  1.46e6
1.48e6
# … with 10,979 more rows, and abbreviated variable names ¹weighed_F,
#   ²weighed_M
```

```
dmf <-dmf |>
mutate(weighed_rate_F=Female.x*Female.y/total_F,weighed_rate_M=Male.x*Male.y/
total_M)

summary_mean <- dmf |> group_by(Year) |>
  summarize(mean_mortality_f = mean(weighed_rate_F, na.rm = TRUE),
            mean_mortality_m = mean(weighed_rate_M, na.rm = TRUE))

summary_mean|> ggplot(aes(x = Year))+geom_line(aes(y=mean_mortality_f),colour
="firebrick4")+geom_line(aes(y=mean_mortality_m),colour ="blue")+
  labs(title = "Female&Male ave_weighted mortality rates over time, Ontario",
       y = "Average Weighted Mortality rate") + scale_color_manual("",values
= c("Female"="firebrick4","Male"="blue"))+ theme_bw(base_size = 14)
```



The trend of average weighted mortality rates over time of both gender was going down until 2000 year and after 2000 year,the trend became a little higher and stable.