Lab 02: Data Visualization

To run this lab yourself you should create a blank quarto document and copy and paste the included code over, to then complete the required tasks.

HIV prevalence from WHO

- Estimated HIV prevalence was obtained from the `gapminder` website https://www.gapminder.org/data/
  - Estimated number of people living with HIV per 100 population of age group 15-49.
  - Original data source is the UNAIDS online database at http://www.aidsinfoonline.org
- A spreadsheet of the data, HIVprev.csv, is necessary for this lab.

We can read in these data as follows (we'll learn about reading in data later in STAT 260):

```
library(tidyverse)

# you must have already installed the tidyverse package

hiv <- read.csv("HIVprev.csv", stringsAsFactors = FALSE)
hiv <- select(hiv, Country, year, prevalence)
```

Take a look at the top and bottom few lines of raw data.

```
head(hiv)
```

```
  Country year prevalence
1 Algeria 1990       0.06
2 Algeria 1991       0.06
3 Algeria 1992       0.06
4 Algeria 1993       0.06
5 Algeria 1994       0.06
6 Algeria 1995       0.06
```

```r
tail(hiv)
```

```
      Country year prevalence
1601 Zimbabwe 1995       25.1
1602 Zimbabwe 1996       26.2
1603 Zimbabwe 1997       26.5
1604 Zimbabwe 1998       26.3
1605 Zimbabwe 1999       25.7
1606 Zimbabwe 2000       24.8
```

```r
summary(hiv)
```

```
   Country               year        prevalence    
 Length:1606        Min.   :1990   Min.   : 0.060  
 Class :character   1st Qu.:1992   1st Qu.: 0.060  
 Mode  :character   Median :1995   Median : 0.200  
                    Mean   :1995   Mean   : 1.575  
                    3rd Qu.:1998   3rd Qu.: 1.100  
                    Max.   :2000   Max.   :26.500  
```

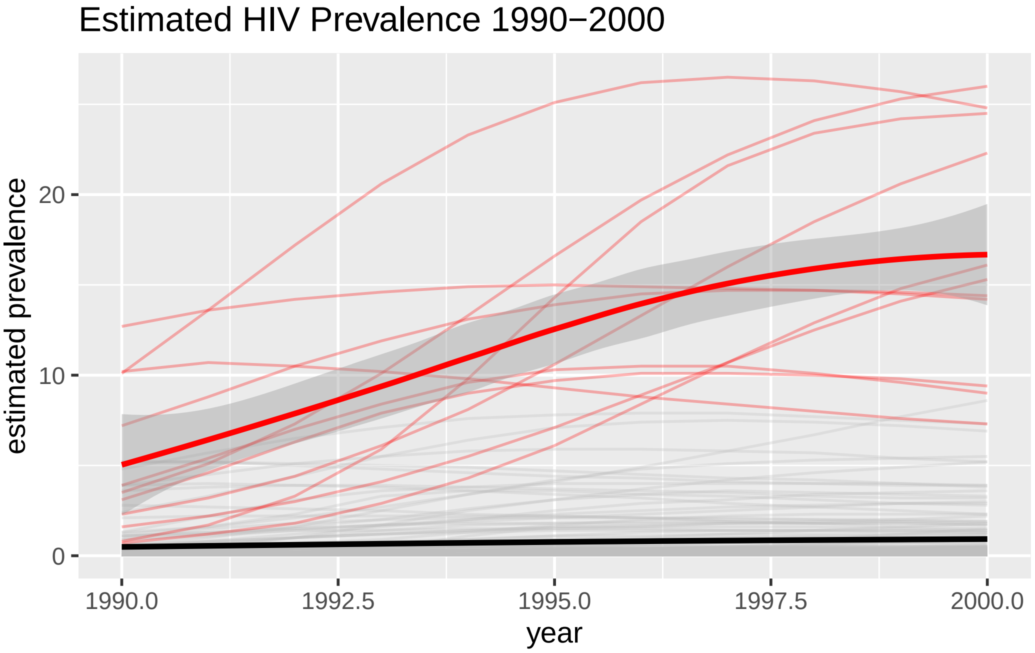
Exercises:

1. Plot the time series of HIV prevalence by year for each country using `geom_line()`.
2. Redo the above plot but experiment with different `alpha` values. What problem does setting a small `alpha` overcome? What feature of the graph is hidden when we do not set `alpha`?
3. In the following code chunk we create a new dataset comprised of countries that had HIV prevalence greater than 10% in one or more of the years monitored (we will learn about this kind of "data wrangling" in future lectures of STAT 260).

```r
cc <- c(
  "Botswana", "Central African Republic", "Congo", "Kenya", "Lesotho", "Malawi",
  "Namibia", "South Africa", "Swaziland", "Uganda", "Zambia", "Zimbabwe"
)
hihiv <- filter(hiv, Country %in% cc)
```

Add red lines for the above countries to your time series plot from Exercise 2.

4. Redo the time series plot from Exercise 1, with the following modifications. Color the time series for all but the countries in the `hihiv` data frame (i.e., those with high HIV prevalence) `grey` and with `alpha=0.3`. For the high-HIV-prevalence countries, color them `red`, also using `alpha=0.3`. Next, add two smoothers: (i) for all the data, i.e. all the countries in the `hiv` data frame, colored `black`, and (ii) for the countries with a high prevalence of HIV, i.e. those in the `hihiv` data frame, colored `red`. Your final plot should look like this (do not worry about axis labels or title):


Estimated HIV Prevalence 1990−2000
estimated prevalence
year