Lab 6

David Stenning

HIV prevalence from WHO

- We used a tidy version of the HIV prevalence data in lab 2, and saw the raw version in lab 3. In this lab we will tidy the latter into the former.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
hiv <- read_csv("HIVprevRaw.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   `Estimated HIV Prevalence% - (Ages 15-49)` = col_character(),
##   `1988` = col_logical(),
##   `1989` = col_logical()
## )
```

```
## See spec(...) for full column specifications.
```

(The columns for 1988 and 1989 are completely empty and were read in as logical. We will be removing these and so won't worry about over-riding the logical with double.)

1. The first column of the data frame is the country, but it has been named`Estimated HIV Prevalence% - (Ages 15-49)`. Use the `rename()` function to rename this column `Country`. (Hint: The current variable name contains special characters and will need to be enclosed in quotes.)
2. The data from 1979 to 1989 is very sparse. Remove these columns from the data frame.
3. Gather the yearly prevalence estimates into key, value pairs with `year` as the key and `prevalence` as the value. When you gather, remove explicity missing values. After gathering, sort on "Country".