Lab 6

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.
```

```
hiv
```

```
## # A tibble: 274 x 34
##    `Estimated HIV ~  `1979` `1980` `1981` `1982` `1983` `1984` `1985` `1986`
##    <chr>               <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Abkhazia         NA           NA     NA     NA     NA     NA     NA     NA
##  2 Afghanistan      NA           NA     NA     NA     NA     NA     NA     NA
##  3 Akrotiri and Dh~ NA           NA     NA     NA     NA     NA     NA     NA
##  4 Albania          NA           NA     NA     NA     NA     NA     NA     NA
##  5 Algeria          NA           NA     NA     NA     NA     NA     NA     NA
##  6 American Samoa   NA           NA     NA     NA     NA     NA     NA     NA
##  7 Andorra          NA           NA     NA     NA     NA     NA     NA     NA
##  8 Angola            0.0265      NA     NA     NA     NA     NA     NA     NA
##  9 Anguilla         NA           NA     NA     NA     NA     NA     NA     NA
## 10 Antigua and Bar~ NA           NA     NA     NA     NA     NA     NA     NA
## # ... with 264 more rows, and 25 more variables: `1987` <dbl>, `1988` <lgl>,
## #   `1989` <lgl>, `1990` <dbl>, `1991` <dbl>, `1992` <dbl>, `1993` <dbl>,
## #   `1994` <dbl>, `1995` <dbl>, `1996` <dbl>, `1997` <dbl>, `1998` <dbl>,
## #   `1999` <dbl>, `2000` <dbl>, `2001` <dbl>, `2002` <dbl>, `2003` <dbl>,
## #   `2004` <dbl>, `2005` <dbl>, `2006` <dbl>, `2007` <dbl>, `2008` <dbl>,
## #   `2009` <dbl>, `2010` <dbl>, `2011` <dbl>
```

(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.)

```
hiv <- rename(hiv,Country = `Estimated HIV Prevalence% - (Ages 15-49)`)
hiv
```

```
## # A tibble: 274 x 34
##    Country  `1979` `1980` `1981` `1982` `1983` `1984` `1985` `1986` `1987`
##    <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Abkhaz~ NA          NA     NA     NA     NA     NA     NA     NA     NA
##  2 Afghan~ NA          NA     NA     NA     NA     NA     NA     NA     NA
##  3 Akroti~ NA          NA     NA     NA     NA     NA     NA     NA     NA
##  4 Albania NA          NA     NA     NA     NA     NA     NA     NA     NA
##  5 Algeria NA          NA     NA     NA     NA     NA     NA     NA     NA
##  6 Americ~ NA          NA     NA     NA     NA     NA     NA     NA     NA
##  7 Andorra NA          NA     NA     NA     NA     NA     NA     NA     NA
##  8 Angola   0.0265     NA     NA     NA     NA     NA     NA     NA     NA
##  9 Anguil~ NA          NA     NA     NA     NA     NA     NA     NA     NA
## 10 Antigu~ NA          NA     NA     NA     NA     NA     NA     NA     NA
## # ... with 264 more rows, and 24 more variables: `1988` <lgl>, `1989` <lgl>,
## #   `1990` <dbl>, `1991` <dbl>, `1992` <dbl>, `1993` <dbl>, `1994` <dbl>,
## #   `1995` <dbl>, `1996` <dbl>, `1997` <dbl>, `1998` <dbl>, `1999` <dbl>,
## #   `2000` <dbl>, `2001` <dbl>, `2002` <dbl>, `2003` <dbl>, `2004` <dbl>,
## #   `2005` <dbl>, `2006` <dbl>, `2007` <dbl>, `2008` <dbl>, `2009` <dbl>,
## #   `2010` <dbl>, `2011` <dbl>
```

2. The data from 1979 to 1989 is very sparse. Remove these columns from the data frame.

```
hiv <- hiv %>% select(-(`1979`:`1989`))
hiv
```

```
## # A tibble: 274 x 23
##    Country `1990` `1991` `1992` `1993` `1994` `1995` `1996` `1997` `1998` `1999`
##    <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Abkhaz~  NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  2 Afghan~  NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  3 Akroti~  NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  4 Albania  NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  5 Algeria   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06   0.06
##  6 Americ~  NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  7 Andorra  NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
##  8 Angola    0.5    0.8    1      1.2    1.4    1.6    1.7    1.8    1.8    1.9
##  9 Anguil~  NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
## 10 Antigu~  NA     NA     NA     NA     NA     NA     NA     NA     NA     NA
## # ... with 264 more rows, and 12 more variables: `2000` <dbl>, `2001` <dbl>,
## #   `2002` <dbl>, `2003` <dbl>, `2004` <dbl>, `2005` <dbl>, `2006` <dbl>,
## #   `2007` <dbl>, `2008` <dbl>, `2009` <dbl>, `2010` <dbl>, `2011` <dbl>
```

3. Pivot the yearly prevalence estimates into a longer tibble that contains only three columns: `Country`, `year`, and `prevalence`. When you pivot, remove explicitly missing values. After pivoting, sort the

resulting tibble by `Country`.

```
hiv %>%
  pivot_longer(c(`1990`:`2011`),
                   names_to="year",
                   values_to="prevalence",
                   values_drop_na=TRUE) %>%
  arrange(Country)
```

```
## # A tibble: 3,212 x 3
##    Country     year  prevalence
##    <chr>       <chr>      <dbl>
##  1 Afghanistan 2009        0.06
##  2 Afghanistan 2010        0.06
##  3 Afghanistan 2011        0.06
##  4 Algeria     1990        0.06
##  5 Algeria     1991        0.06
##  6 Algeria     1992        0.06
##  7 Algeria     1993        0.06
##  8 Algeria     1994        0.06
##  9 Algeria     1995        0.06
## 10 Algeria     1996        0.06
## # ... with 3,202 more rows
```