Stat 261, Lab 4

David Stenning

HIV prevalence from WHO

- The HIV prevalence data from lab 2 was modified from its raw form. In this lab we will work with the raw data to explore data transformation.
- A spreadsheet of the data `HIVprevRaw.csv` is available in the Labs–>Lab4 folder on Canvas. We can read in these data as follows (we'll learn about reading in data in later weeks):

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- 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
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

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

```
hiv <- read.csv("HIVprevRaw.csv",stringsAsFactors = FALSE)
```

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

```
head(hiv)
```

```
##   Estimated.HIV.Prevalence.....Ages.15.49. X1979 X1980 X1981 X1982 X1983 X1984
## 1                                 Abkhazia    NA    NA    NA    NA    NA    NA
## 2                              Afghanistan    NA    NA    NA    NA    NA    NA
## 3                    Akrotiri and Dhekelia    NA    NA    NA    NA    NA    NA
## 4                                  Albania    NA    NA    NA    NA    NA    NA
## 5                                  Algeria    NA    NA    NA    NA    NA    NA
## 6                           American Samoa    NA    NA    NA    NA    NA    NA
##   X1985 X1986 X1987 X1988 X1989 X1990 X1991 X1992 X1993 X1994 X1995 X1996 X1997
## 1    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 2    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 3    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 4    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 5    NA    NA    NA    NA    NA  0.06  0.06  0.06  0.06  0.06  0.06  0.06  0.06
```

```
## 6    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
##   X1998 X1999 X2000 X2001 X2002 X2003 X2004 X2005 X2006 X2007 X2008 X2009 X2010
## 1    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 2    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA  0.06  0.06
## 3    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 4    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 5  0.06  0.06  0.06  0.06  0.06  0.06   0.1   0.1   0.1   0.1   0.1    NA    NA
## 6    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
##   X2011
## 1    NA
## 2  0.06
## 3    NA
## 4    NA
## 5    NA
## 6    NA
```

- Make a copy of `hiv` for use in exercise 4.

```
hivcopy <- hiv
```

- In exercises 1 - 3, save the results of each data manipulation in `hiv`. In exercise 4 you will use the copy `hivcopy`.

Exercises:

1. The first column of the data frame is the country, but it has been named: `Estimated.HIV.Prevalence.....Ages.15.49.1.` Use the `rename()` function to rename this column `Country`.
2. The data from 1979 to 1989 is very sparse. Remove these columns from the data frame.
3. Sort the data in descending order of prevalence in 2011. Print the first 6 rows of your final data set.
4. Use the copy `hivcopy` and the pipe operator to chain or "pipe" the data manipulations of exercises 1-3.