Data Jujutsu II – PhD Trends*

Stefano Allesina & Graham Smith *University of Chicago*

Description of the data

Every year, the National Science Foundation sponsors a very large survey (with almost complete sampling) of the PhD graduates, the *Survey of Earned Doctorates* (SED). They publish statistics on the number of PhD, and report PhD completion by gender, field, ethnic background, etc. In particular, table 16 reports the number of PhD awarded by sex and field of study. We are going to attempt reading the table directly from the `xlsx` files that are published by NSF.

The challenge

1. The file `urls_and_skip_NSF_SED.csv` reports the location (`url`) of the excel files for the years 2013-1018, as well as the number of lines to skip (`skip`) and the number of lines to read (`read`) for best results. Read the documentation of `read_xlsx` from the library `readxl` to see how to read the file while skipping a few lines and capping the total number of lines to be read.

```
library(tidyverse)
library(readxl)
read_csv("urls_and_skip_NSF_SED.csv")
```

```
## # A tibble: 6 x 4
##    year url                                                      skip  read
##   <dbl> <chr>                                                  <dbl> <dbl>
## 1  2018 https://ncses.nsf.gov/pubs/nsf20301/assets/data-tables/tabl~     3   274
## 2  2017 https://ncses.nsf.gov/pubs/nsf19301/assets/data/tables/sed1~     3   271
## 3  2016 https://nsf.gov/statistics/2018/nsf18304/data/tab16.xlsx         1   270
## 4  2015 https://nsf.gov/statistics/2017/nsf17306/data/tab16.xlsx         1   264
## 5  2014 https://nsf.gov/statistics/2016/nsf16300/data/tab16.xlsx         1   293
## 6  2013 https://nsf.gov/statistics/sed/2013/data/tab16.xlsx              1   284
```

Read all the files, building the tibble `sed` with structure:

```
source("solution_PhD_trends.R") # this is the code you have to write!
sed
```

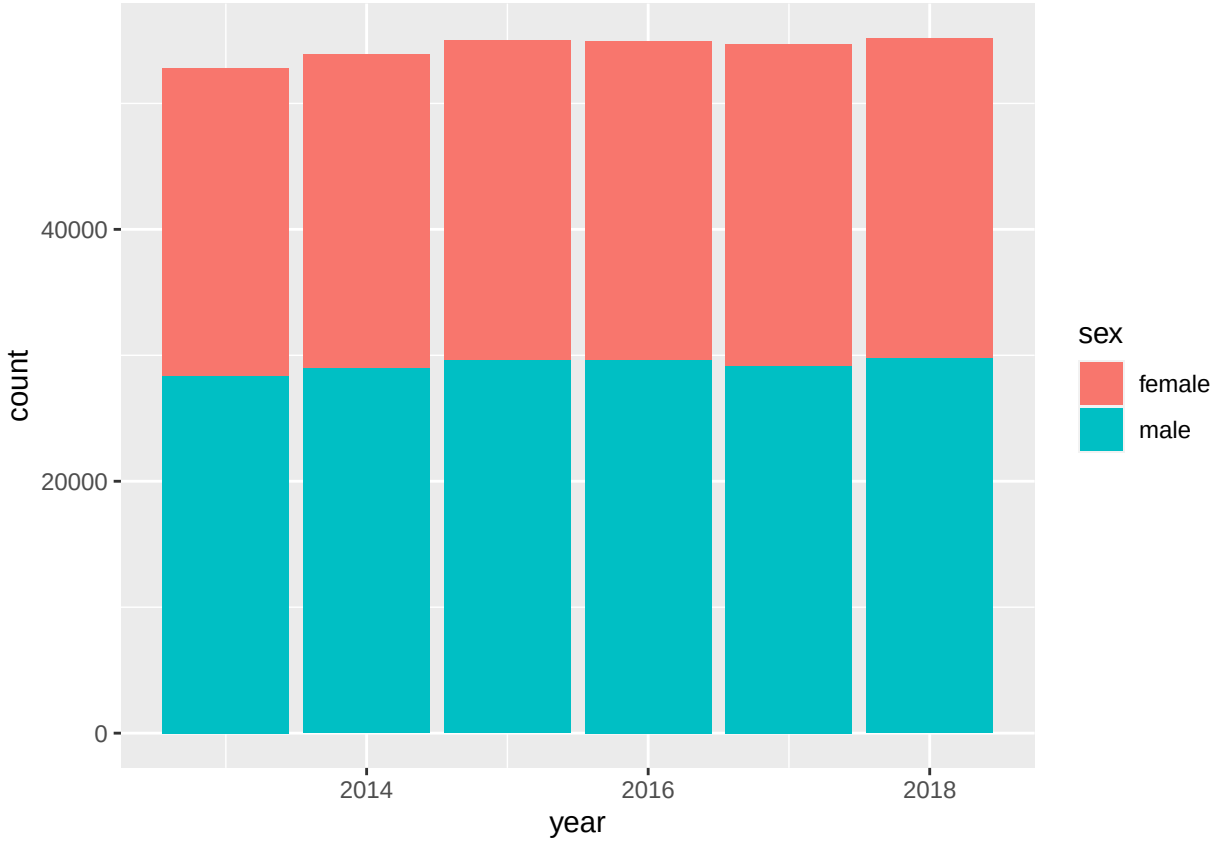
```
## # A tibble: 1,583 x 4
##    field                                                  male female  year
##    <chr>                                                 <dbl>  <dbl> <dbl>
##  1 All fields                                            29798  25368  2018
##  2 Life sciences                                          5659   7114  2018
```

*This document is included as part of the Advanced Computing I packet for the U Chicago BSD qBio6 boot camp 2020. **Current version**: July 06, 2020; **Corresponding author**: sallesina@uchicago.edu.

```
##  3 Agricultural sciences and natural resources                746    696  2018
##  4 Agricultural sciences                                      458    416  2018
##  5 Agricultural economics                                      65     43  2018
##  6 Agronomy, horticulture science, plant breeding, plant pat~  209    140  2018
##  7 Animal nutrition, poultry science                           36     31  2018
##  8 Animal sciences, other                                      48     73  2018
##  9 Food science, food technology-other                         67     96  2018
## 10 Soil chemistry and microbiology, soil sciences-other        33     33  2018
## # ... with 1,573 more rows
```

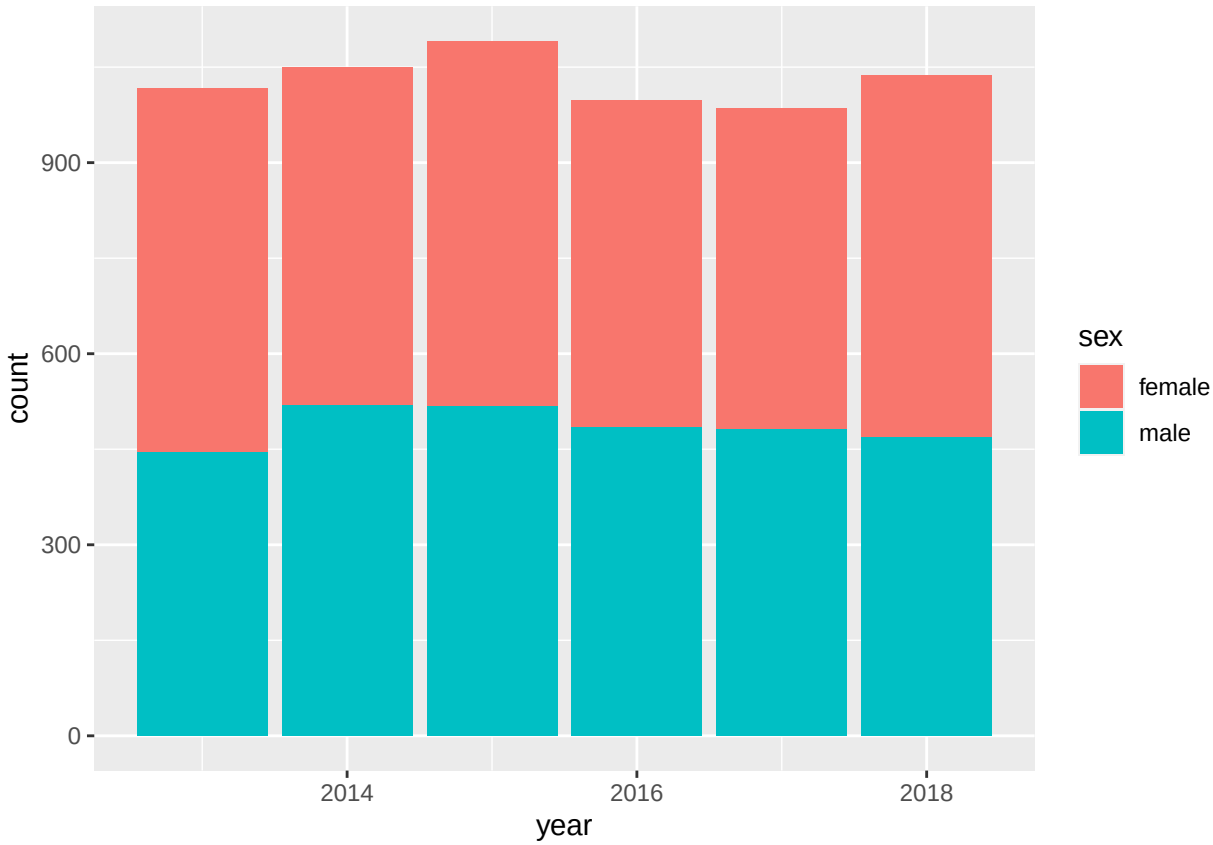
2. Write a generic function for plotting, and plot the number of PhDs in time, by taking the `All fields`:

```
plot_PhD_in_time(sed, "All fields")
```



Produce the same type of graph for the disciplines that interest you. Note that the naming of some of the fields has changed: for example, you find `Neurosciences, neurobiology` (recent years) and `Neurosciences and neurobiology` (older data sets). Modify the function such that it uses `grepl` to match a given label (`Neurosciences` in this case).

```
plot_PhD_in_time(sed, "Neurosciences")
```



3. The graduates in some of the fields are predominantly male (e.g., `Robotics`), while in other fields most graduates are females (e.g., `Developmental and child psychology`). Find the biological field having the largest gender disparity.
4. [Optional] Find the biological field that has seen the greatest change in gender composition in time.

Hints & Nifty tricks

- If you don't want to store the downloaded zip file, use a temporary file (it will be deleted by `R` automatically once you call `unlink()`)
- Some lines are empty: use something like `filter(!is.na(field))` to get rid of them.
- For each year, you only need to store the number of PhD awarded to men/women; the rest of the information is redundant, and can be calculated from these two numbers.