INTRODUCTION TO DATA SCIENCE WITH R

Session 6a: Importing data

Ina Bornkessel-Schlesewsky

January 20, 2023

WORKING WITH DATA FROM A FILE

- Typically, the data we work with aren't available as R packages
- Rather, we need to read them in from a file
- csv is a popular format
  - comma separated values
  - plain text
  - can be exported from all popular spreadsheet applications (e.g. Excel)
  - (+ is accessible without any proprietary software, allows for version control etc. – more on this later …)

EXAMPLE

Student-to-teacher ratios in different parts of the world:

- data from the UNESCO Institute for statistics
- made available via the Tidy Tuesday challenge
- preprocessed by Cédric Scherer – see blogpost

DATA SET

```
st_ratios <- read_csv("student_teacher_ratios.csv")
glimpse(st_ratios)
```

```
Rows: 180
Columns: 20
$ indicator            <chr> "Primary Education", "Primary Education", "Primar…
$ country              <chr> "Afghanistan", "Albania", "Algeria", "Angola", "A…
$ country_code         <chr> "AFG", "ALB", "DZA", "AGO", "ATG", "ARG", "ARM", …
$ edulit_ind           <chr> "PTRHC_1", "PTRHC_1", "PTRHC_1", "PTRHC_1", "PTRH…
$ year                 <dbl> 2017, 2017, 2017, 2015, 2017, NA, NA, 2017, 2017,…
$ student_ratio        <dbl> 44.00995, 17.94478, 24.22505, 50.02951, 12.05576,…
$ flag_codes           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ flags                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ name                 <chr> "Afghanistan", "Albania", "Algeria", "Angola", "A…
$ alpha.2              <chr> "AF", "AL", "DZ", "AO", "AG", "AR", "AM", "AT", "…
$ alpha.3              <chr> "AFG", "ALB", "DZA", "AGO", "ATG", "ARG", "ARM", …
$ country.code         <chr> "004", "008", "012", "024", "028", "032", "051", …
$ iso_3166.2           <chr> "ISO 3166-2:AF", "ISO 3166-2:AL", "ISO 3166-2:DZ"…
$ region               <chr> "Asia", "Europe", "Africa", "Africa", "North Amer…
$ sub.region           <chr> "Southern Asia", "Southern Europe", "Northern Afr…
$ region.code          <chr> "142", "150", "002", "002", "019", "019", "142", …
$ sub.region.code      <chr> "034", "039", "015", "017", "029", "005", "145", …
$ x                    <dbl> 22, 15, 13, 13, 7, 6, 20, 15, 21, 4, 20, 23, 8, 1…
$ y                    <dbl> 8, 9, 11, 17, 4, 14, 6, 6, 7, 2, 9, 8, 6, 4, 5, 3…
$ student_ratio_region <dbl> 19.64278, 13.01069, 36.38758, 36.38758, 16.18269,…
```

TO WORK WITH THIS DATA SET

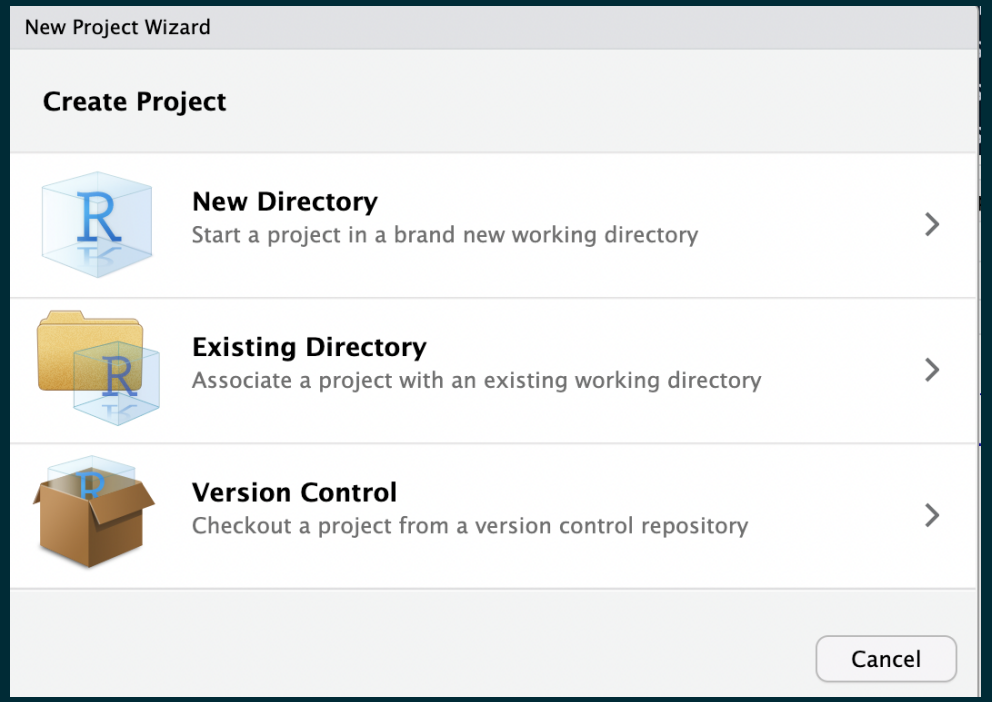
- Download the zip-archive from the course website (under Schedule & Materials > Resources)
- Move the file *student_teacher_ratios.csv* to a suitable directory on your computer
- Create a new RStudio project (see next slide)

A PROJECT-BASED WORKFLOW

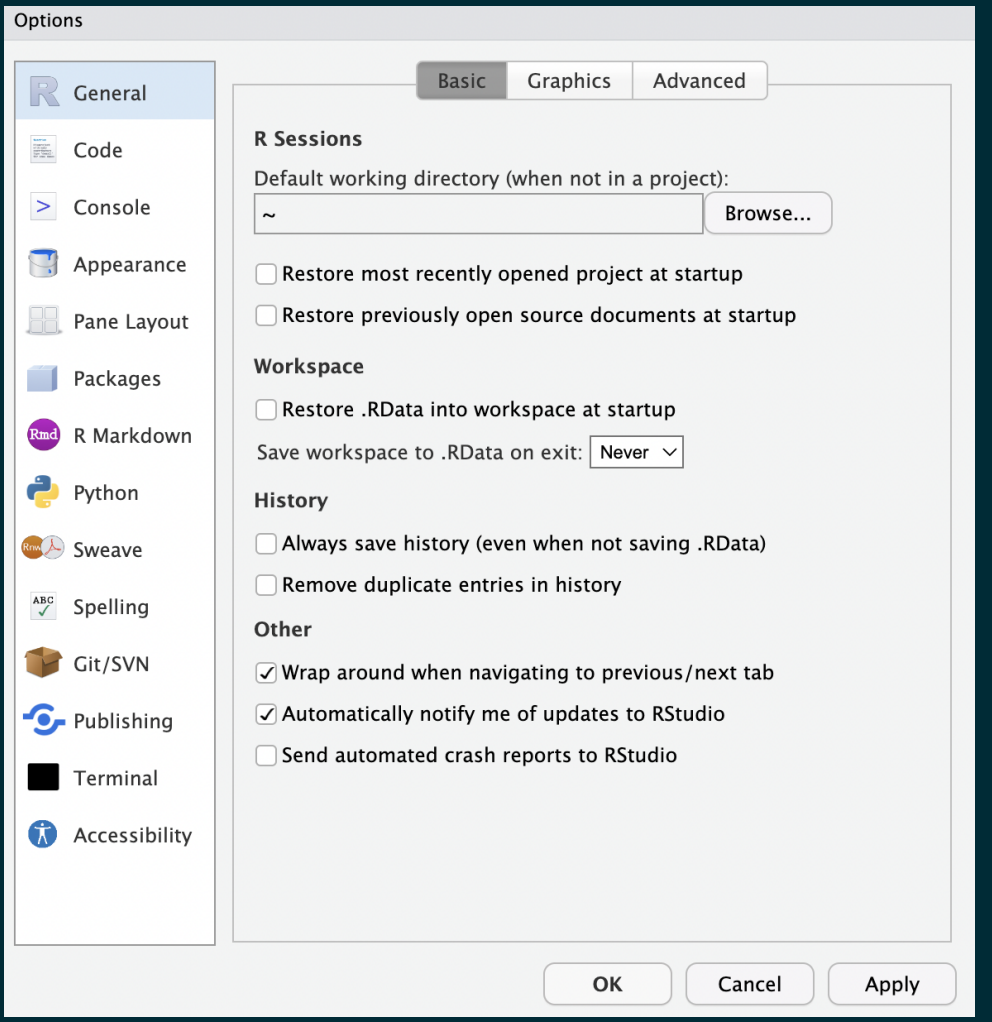
- RStudio projects are a great way to keep all associated components of a project in the one place:
  - Analysis code (R scripts / Quarto documents)
  - Data
  - Results
- This further enhances the reproducibility of your analysis

CREATING AN RSTUDIO PROJECT

- File > New Project
- choose “Existing directory” – the one that you saved the .csv file to
- (as you can see, there are also other options)



- Check RStudio settings



CREATING AN RSTUDIO PROJECT

Suggested reading

R for Data Science - Chapter 9 provides further details on and motivation for a project-based workflow.

READING IN THE FILE

- open your RStudio project by double clicking on the .Rproj file
- this will ensure that R's "working directory" is set correctly
- make sure that both your Quarto file and the csv file are both in the top-level project directory
  - for more complex directory structures, check out the {here} package

- use the `read_csv()` function from the {readr} package (part of the `tidyverse`)
- you will need to create a new object

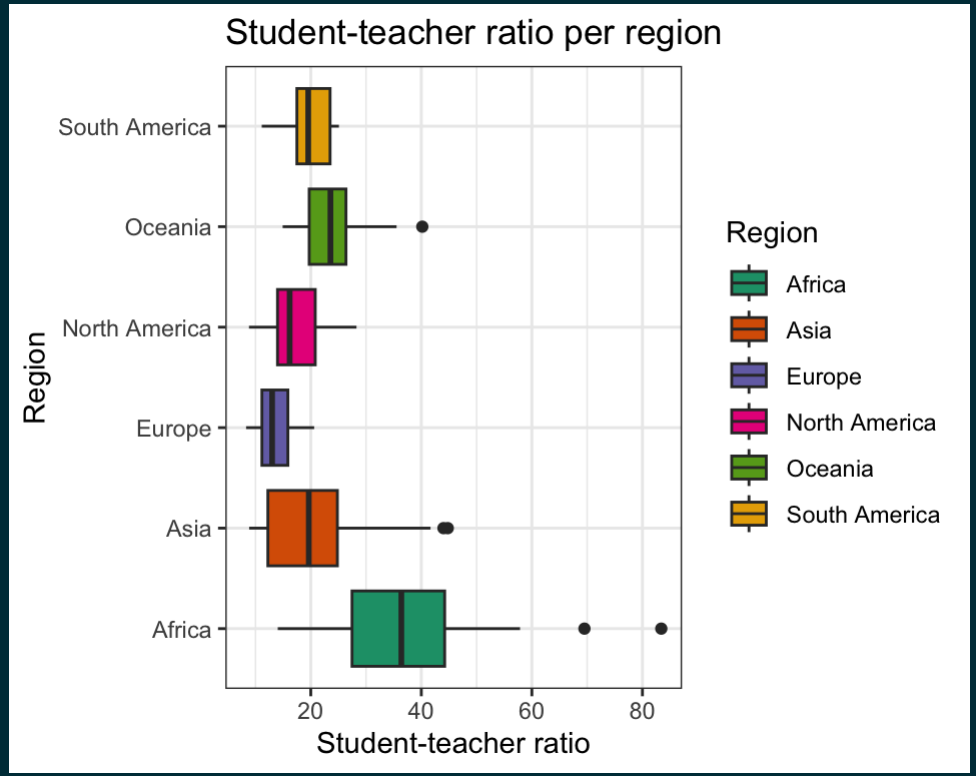
```
st_ratios <- read_csv("student_teacher_ratios.csv")
```

BRIEF EXPLORATION OF STUDENT-TEACHER RATIOS

JOINT EXPLORATION

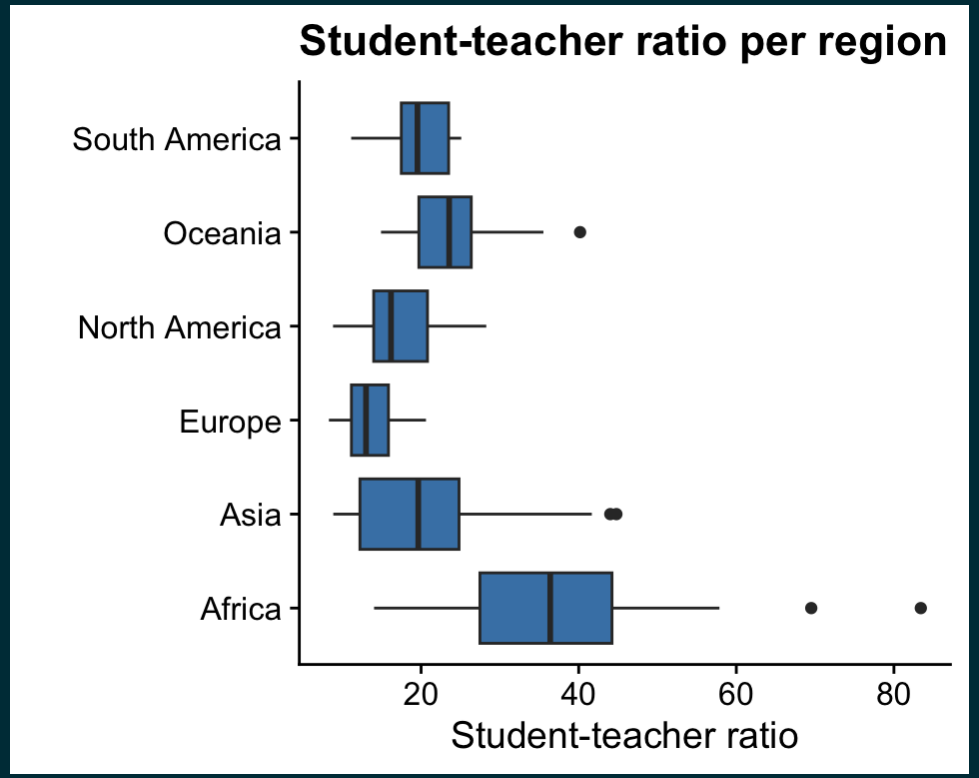
1. Which region has the highest variability in student-teacher ratios? Visualise this with an appropriate figure. *Note the use of flipped x/y axes as well as a custom colour palette and theme in the example.*

```
st_ratios |>
  ggplot(aes(x = region,
             y = student_ratio,
             fill = region)) +
  geom_boxplot() +
  # use a custom colour scale
  scale_fill_brewer(palette = "Dark2") +
  # use a custom theme
  theme_bw() +
  # flip x and y axes for a horizontal boxplot
  coord_flip() +
  labs(
    title = "Student-teacher ratio per region",
    x = "Region",
    y = "Student-teacher ratio",
    fill = "Region"
  )
```



ALTERNATIVE VERSION

```
st_ratios |>
  ggplot(aes(x = region,
             y = student_ratio)) +
  # use a custom colour to fill all pi
  # (not an aesthetic)
  geom_boxplot(fill = "steelblue") +
  # use another custom theme
  # for this one, we need to install
  # the {cowplot} package
  theme_cowplot() +
  # flip x and y axes for a horizontal
  coord_flip() +
  labs(
    title = "Student-teacher ratio per
    x = "",
    y = "Student-teacher ratio",
    fill = "Region"
  )
```



A FEW QUESTIONS TO EXPLORE

2. Focus on the region that you identified in 1.

- Create a new dataframe just for this region.
- Which country has the lowest ST-ratio in this region and which has the highest? What are these?
- Isolate the countries with a ST-ratio higher than the median and plot the ST-ratios for these using an ordered horizontal column graph (recall from the week 4 exercises how to change the order). Use an appealing colour for the columns and pick a theme for the plot that you like.

(If you would like extra practice: try doing the same as in 2 for the region with the *lowest* variability.)

RESOURCES

- to see list of in-built colours in R, use `colours()`
- for information on available colour palettes in R, see https://github.com/EmilHvitfeldt/r-color-palettes (note that some of these require the installation of new packages – more on this later)
- list of ggplot themes: https://ggplot2.tidyverse.org/reference/ggtheme.html (again, more themes are available via other packages, e.g. a theme to produce APA-style plots)