Chapter 2 HW 2. First data skills

2.1 Learning Outcomes

By the end of this chapter, you should be able to:

- Install and load R packages
- Load a dataset and see its contents
- Explain the types of data that R uses, and the types of variables that can be captured.

2.2 Activity 1: Setup

We can work in the same project as in the Intro to R and RStudio HW.

- Open your R project by double-clicking it, or from within RStudio via the **File > Open Project…** menu.
- Open a new R Markdown notebook: click **File > New File > R Notebook** or click on the little page icon with a green plus sign (top left).
- Give it a meaningful `title` (e.g., 'HW 2: Data Skills') - you can also change the title later. Feel free to add an `author` field with your name or Andrew ID.
- Once the `.Rmd` is opened, you need to save the file.
- To save it, click **File > Save As…** or click on the little disc icon. Name it something meaningful (e.g., "data-skills-hw2.Rmd"). Make sure there are no spaces in the name - R is not very fond of spaces… This file will automatically be saved in your project folder (i.e., your working directory) so you should now see this file appear in your file viewer pane.

Note: Don't ever save a new project **inside** another project directory. This can cause some hard-to-resolve problems.

2.3 Activity 2: Download the data

The data we will use for this hw is adapted from a paper by Pownall et al.

> Pownall, M., Pennington, C. R., Norris, E., Juanchich, M., Smailes, D., Russell, S., Gooch, D., Evans, T. R., Persson, S., Mak, M. H. C., Tzavella, L., Monk, R., Gough, T., Benwell, C. S. Y., Elsherif, M., Farran, E., Gallagher-Mitchell, T., Kendrick, L. T., Bahnmueller, J., . . . Clark, K. (2023). Evaluating the Pedagogical Effectiveness of Study Preregistration in the Undergraduate Dissertation. *Advances in Methods and Practices in Psychological Science, 6*(4). https://doi.org/10.1177/25152459231202724

Download it here: data_ch1.zip. There are 2 csv files contained in a zip folder. One is the data file we are going to use today `prp_data_reduced.csv` and the other is an Excel file `prp_codebook` that explains the variables in the data.

The first step is to **unzip the zip folder** so that the files are placed within the same folder as your project.

- Place the zip folder within your Data-Skills-HW folder
- Right mouse click –> `Extract All...`
- Check the folder location is the one to extract the files to
- Check the extracted files are placed next to the project icon
- Files and project should be visible in the Output pane in RStudio

2.4 Activity 3: Installing packages, loading packages, and reading in data

2.4.1 Installing packages

When you install R and RStudio for the first time (or after an update), most of the packages we will be using won't be pre-installed. Before you can load new packages like `tidyverse`, you will need to install them.

If you try to load a package that has not been installed yet, you will receive an error message that looks something like this: `Error in library(tidyverse) : there is no package called 'tidyverse'`.

To fix this, simply install the package first. **In the console**, type the command `install.packages("tidyverse")`. This **only needs to be done once after a fresh installation**. After that, you will be able to load the `tidyverse` package into your library whenever you open RStudio.

Note: Never include `install.packages()` in the Rmd. Only install packages from the console pane or the packages tab of the lower right pane!!!

There will be other packages used in later chapters that will also need to be installed before their first use, so this error is not limited to `tidyverse`.

2.4.2 Loading packages and reading in data

The first step is to load in the packages we need and read in the data. Today, we'll only be using `tidyverse`, and `read_csv()` will help us store the data from `prp_data_reduced.csv` in an object called data_prp.

Copy the code into a code chunk in your `.Rmd` file and run it. You can either click the `green arrow` to run the entire code chunk, or use the shortcut `Ctrl + Enter` (Windows) or `Cmd + Enter` (Mac) to run a line of code/ pipe from the Rmd.

```r
library(tidyverse)
data_prp <- read_csv("prp_data_reduced.csv")
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (http://conflicted.r-lib.org/) to force all conflicts

Rows: 89 Columns: 91
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (17): Code, Age, Ethnicity, Opptional_mod_1_TEXT, Research_exp_1_TEXT, U...
dbl (74): Gender, Secondyeargrade, Opptional_mod, Research_exp, Plan_prereg,...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

2.5 Activity 4: Familiarize yourself with the data

- Look at the **Codebook** to get a feel of the variables in the dataset and how they have been measured. Note that some of the columns were deleted in the dataset you have been given.
- You'll notice that some questionnaire data was collected at 2 different time points (i.e., SATS28, QRPs, Understanding_OS)
- Some of the data was only collected at one time point (i.e., supervisor judgements, OS_behav items, and Included_prereg variables are t2-only variables)

2.5.1 First glimpse at the data

Before you start wrangling your data, it is important to understand what kind of data you're working with and what the format of your dataframe looks like.

As you may have noticed, `read_csv()` provides a **message** listing the data types in your dataset and how many columns are of each type. Plus, it shows a few examples columns for each data type.

To obtain more detailed information about your data, you have several options. Test them out in your own `.Rmd` file and use whichever method you prefer (but do it).

- ▶ **Visual inspection (approach 1)**
- ▶ **Visual inspection (approach 2)**
- ▶ **`glimpse()`**
- ▶ **`spec()`**

2.5.2 Your Turn

Now that you have tested out all the options in your own R Notebook, you can probably answer the following questions:

- How many observations?
- How many variables?
- How many columns are `col_character` or `chr` data type?
- How many columns are `col_double` or `dbl` data type?

In your `.Rmd`, include a **new heading level 2** called "Information about the data" (or something equally meaningful) and jot down some notes about `data_prp`. You could include the citation and/or the abstract, and whatever information you think you should note about this dataset (e.g., any observations from looking at the codebook?). You could also include some notes on the functions used so far and what they do. Try to incorporate some **bold**, *italic* or ***bold and italic*** emphasis and perhaps a bullet point or two.

2.5.3 Data types

Each variable has a **data type**, such as numeric (numbers), character (text), and logical (TRUE/FALSE values), or a special class of factor. As you have just seen, our `data_prp` only has character and numeric columns (so far).

Numeric data can be double (`dbl`) or integer (`int`). Doubles can have decimal places (e.g., 1.1). Integers are the whole numbers (e.g., 1, 2, -1) and are displayed with the suffix L (e.g., 1L). This is not overly important but might leave you less puzzled the next time you see an L after a number.

Characters (also called "strings") is anything written between quotation marks. This is usually text, but in special circumstances, a number can be a character if it placed within quotation marks. This can happen when you are recoding variables. It might not be too obvious at the time, but you won't be able to calculate anything if the number is a character.

You can use the `typeof()` command to check what data type a variable is.

```
typeof(1)
typeof(1L)
typeof("1")
typeof("text")
```

```
[1] "double"
[1] "integer"
[1] "character"
[1] "character"
```

Logical data (also sometimes called "Boolean" values) are one of two values: TRUE or FALSE (written in uppercase). They become really important when we use `filter()` or `mutate()` with conditional statements such as `case_when()` , as we will see shortly.

A **factor** is a specific type of integer or character that lets you assign the order of the categories. This becomes useful when you want to display certain categories in "the correct order" either in a dataframe (see *arrange*) or when plotting. More on this later.

2.5.4 Variable types

You've already encountered this idea, but let's refresh. Variables can be classified as **continuous** (numbers) or **categorical** (labels).

Categorical variables are properties you can count. They can be **nominal**, where the categories don't have an order (e.g., gender) or **ordinal** (e.g., Likert scales either with numeric values 1-7 or with character labels such as "agree", "neither agree nor disagree", "disagree"). Categorical data may also be **factors** rather than characters.

Continuous variables are properties you can measure and calculate sums/ means/ etc. They may be rounded to the nearest whole number, but it should make sense to have a value between them. Continuous variables always have a **numeric** data type (i.e. `integer` or `double`).

2.6 Activity 5: Save and quit!

We're done with all the coding activities for this HW, so save and quit.

- First, make sure your R Markdown is saved. If there are any unsaved changes then the save icon will be in blue, if it's greyed out it means there are no unsaved changes. But just to be safe, always hit `Ctrl + s` or click `File – Save` which will save your file.
- Then, exit RStudio completely by clicking "File - Quit session" or using the shortcut "Ctrl + Q" (Windows) or "Cmd + Q" (Mac).

2.7 Check for completeness

For this HW, upload your R notebook (.Rmd file), and your Preview html file (.nb.html). Both should include:

- Code chunks that load the tidyverse packages, read in the data, and inspect it.
- Your "Information about the data" notes, answering the four questions above (plus any other notes you may find useful)
- Any notes for yourself about things you figured out along the way

All code in your R Notebook should run successfully.