Theme 1: Data Wrangling Essentials - Assignments (16th of March, 2023)

Eemil Mitikka | IQDA Course 2023 | University of Helsinki

Some general points on all course assignments:

- The language of the course is English, so let's try to talk, think and write in English all the time. **Do not worry too much about grammar etc.**
- Reading other students' assignment reports is ***warmly recommended.***
- The idea is to work with the assignments independently, but **the reports will be openly available for other students after the submission deadline**. This way we can **discuss the reports together** and learn more.
- Errors and mistakes are an important part of learning. **Do not waste your time being afraid of making errors** – just do your best. ***It is essential that you try, otherwise you will not learn!***
- **Asking questions is good activity!** You may ask other students too. You can also answer any questions, not just ask. Questions and answers will always help others as well.

How to Read Exercises – General Instructions

- In the weekly exercises, the assignments are color coded with blue, tasks with red and and 🧐, and support code in the help file with dark green and 👩‍💻.

About the first assignment (Theme 1 – Data Wrangling Essentials)

This is the first time I'm organizing this course, so some of the themes, topics, and things might be harder to you than I expected or I might not explain some things clearly enough. The general workflow in R/RStudio seemed to be one of these things based on today's (16th of March) lecture. Hence, I decided that ***if*** you're really struggling with the original assignment I planned for Theme 1, you can just replicate the "Play around material for Theme 1" material (with the gapminder data) that we started to work with during 16th of Marchlecture and submit these materials as Theme 1 Weekly Assignment. ***However, try first to do the original assignment with the Varieties of Democracy (V-Dem) data (assignment given below).*** Don't be afraid of mistakes, we are here to learn new things.

Reminder about R workflow 💻

Since quite many seemed to be wondering why their R code doesn't work, here are some remarks:

1. ALWAYS start with activating the R packages and by importing the data

This rule applies to any R coding – **R simply does not know what it's supposed to do, if your data, functions, and so on don't exist in your R environment**. Thus, start every R Markdown file with the following lines of code:

```
library(tidyverse)
library(here)

# Name your data with <- , use here() and filename (in this case, "yourdata.csv"), %>% operator (which means AND THEN), and read the data with read.yourdataformat() function (in this case: read.csv() because you have a .csv file)
your_data_name <- here("yourdata.csv") %>%
  read.csv()
```

2. BE PRECISE

R code needs to be written precisely. If the name of your R objects (e.g. data frame, vectors, etc.) lives only in your head, R doesn't know that (if it would know, that would be actually quite scary). Therefore, **you need to define every object you want use in R**. Also, ***R doesn't accept typos***, so you need to pay attention that you use precisely the same object names you have defined in R.

Below, you may find an example of how R reacts even to a tiny typo:

```
# Assing value for life_expectancy integer vector
life_expectancy <- 76

# Try to print the this integer vector with a tiny typo (extra "y" in the end)
life_expectancyy
```

```
## Error in eval(expr, envir, enclos): object 'life_expectancyy' not found
```

As you can see, R doesn't print the `life_expectancyy` vector because of the typo and it gives an error:

"## Error in eval(expr, envir, enclos): object 'life_expectancyy' not found".

Hence, you need to be accurate when writing R code.

3. PAY ATTENTION THE ORDER WHEN YOU WRITE CODE

R code needs to be written logically. For example, if you let's assume you have a data frame called `my_data` which consists of `gapminder` data. You want to select only certain variables and print the first rows of the data using `select()` and `head()` functions, and you have loaded the needed *tidyverse* and *here* packages needed to import the data with `library()` function call. You're also assinging the `my_data` R object appropriately with the `<-` operator.

However, in the code below, the definition of your `my_data` object comes **only after** the `my_data %>% select(year, lifeExp, gdpPercap)` call. Hence, R throws an error complaining:

"## Error in select(., year, lifeExp, gdpPercap): object 'my_data' not found"

```
library(tidyverse)
library(here)

# Import the data
gapminder <- here("gapminder.csv") %>%
  read.csv()

# (Try to) select only year, lifeExp, gdpPercap columns, and print the first 6 observations
my_data %>%
  select(year, lifeExp, gdpPercap) %>%
  head()
```

```
## Error in select(., year, lifeExp, gdpPercap): object 'my_data' not found
```

```
# This should be defined before calling
# my_data %>%
#   select(year, lifeExp, gdpPercap)
my_data <- gapminder
```

Thus, R code needs to written in logical order. Below, you may find how the following code should be written:

```
library(tidyverse)
library(here)

# Import the data
gapminder <- here("gapminder.csv") %>%
  read.csv()

# Assign the R object
my_data <- gapminder

# (Try to) select only year, lifeExp, gdpPercap columns
my_data %>%
  select(year, lifeExp, gdpPercap) %>%
  head()
```

```
##   year lifeExp gdpPercap
## 1 1952  28.801  779.4453
## 2 1957  30.332  820.8530
## 3 1962  31.997  853.1007
## 4 1967  34.020  836.1971
## 5 1972  36.088  739.9811
## 6 1977  38.438  786.1134
```

The Original Theme 1 Assignment – Data Wrangling with Varieties of Democracy (V-Dem) data

In the first theme's exercises, we use the Varieties of Democracy (V-Dem) data to explore the state of democracy across the globe. We will work especially with the variable called `v2x_libdem`, which is a index score measuring the level of liberal democracy in different countries. The liberal democracy index varies between 0 and 1, where 0='Lowest level of liberal democracy', and 1='Highest level of liberal democracy'.

For this exercise, I subset the original V-Dem data to include only couple of variables since the original dataset is quite large. However, you may find the original V-Dem dataset from Theme 1's assignments folder on Moodle. You may also download these data for free from the V-Dem's web site. I also added the variable codebook, if you want to take a closer look at the variables and their definitions.

The variables included in this assignment dataset are:

- `X`, which is just the row name of the data
- `year`, which is the year of observed values for variables
- `v2x_libdem`, which is the liberal democracy index
- `v2x_civlib`, which is the civil liberties index
- `v2x_corr`, which is the political corruption index

Below, you may find the "raw" assignments without any support code. Use RStudio and R Markdown in your work, and submit your assignments before the deadline on ***Moodle as .html and .Rmd files***.

Assignment 1

- *How many variables and observations there are in the `vdem` data?* 🧐

Assignment 2

- *Subset the `vdem` data to include only the variables `country_name`, `year`, and `v2x_libdem` and observations from year 2022. Create a new data frame for these data and name it as `vdem_libdem_2022`.* 🧐

Assignment 3

- *Find out which five countries score highest on liberal democracy index (column: `v2x_libdem`) in the `vdem_libdem_2022` data frame.* 🧐
- *Find out the number of unique liberal democracy index scores in `vdem_libdem_2022` data frame.* 🧐

Assignment 4

Let's imagine a following scenario: you're working in a research project dealing with comparing the state of democracy across the world. You have discovered the V-Dem dataset and want to use the liberal democracy index data for year 2022 in the final report of the project. 🧐

Your boss likes the idea, but wants this scale to vary from **0 to 10** *instead of* **0 to 1**, *as in the original V-Dem data because other indices in the final report also vary from 0 to 10. Create a new variable `libdem_0_to_10` that captures the liberal democracy index, but ranging from 0 to 10 instead of 0 to 1.* 🧐

Assignment 5

- *Calculate the mean and median for liberal democracy index in 2022.* 🧐
- *Calculate minimum and maximum years present in the `vdem` data.* 🧐

Assignment 6

- *Calculate mean liberal democracy index, `v2x_libdem`, by countries. Find out which are the top and bottom three countries on liberal democracy index.* 🧐

Assignment 7

- *Find out which six countries have the least observations in the whole `vdem` data.* 🧐