Cleaning and Wrangling Data

Fundamentals of R Workshop - Homework 2

- All materials for the exercises below are available in the homework folder on Moodle.
- Please submit an R script containing both the code and results on Moodle.
- For each question, please do not forget to include the code used to find the answer.
- You can #comment out any sentences that are part of your answers but are not R code.

Ph.D. theses at the Graduate Institute II

The Institute has been the home for hundreds of Ph.D. students over time. In this homework, we will use a dataset on Ph.D. theses at the Institute from last week, but we will augment it with new data.

Question 1

The `topic.xlsx` file contains the extent to which Ph.D. theses address topics related to gender and sustainability, but it is messy. Import the dataset and clean it so it contains the following variables:

Name	Description
`thesis_ID`	ID of the Ph.D. thesis.
`sus_degree`	Extent to which theses covers sustainaibility topics.
`gen_degree`	Extent to which theses covers gender-related topics.

Question 2

The `department.csv` file contains information on the department and language of a thesis, but it is messy. Import the dataset and clean it so it contains the following variables:

Name	Description
`thesis_ID`	ID of the Ph.D. thesis.
`thesis_department`	Department in which the thesis was written.
`thesis_language`	Language in which the thesis was written.

Question 3

The `phd_theses.rds` file contains information on theses' author and year. Merge this dataset to the datasets in question 1 and question 2. The new "phd_theses" dataset should contain the following variables:

Name	Description
`thesis_title`	Title of the Ph.D. thesis.

Name	Description
`thesis_ID`	ID of the Ph.D. thesis.
`thesis_year`	The year in which a thesis was submitted.
`thesis_author`	The author of the thesis.
`thesis_department`	Department in which the thesis was written.
`thesis_language`	Language in which the thesis was written.
`sus_degree`	Extent to which theses covers sustainability topics.
`gen_degree`	Extent to which theses covers gender-related topics.

Question 4

Inspect the dataset. How many departments are there? In how many languages were theses written? When was the first thesis written?

Question 6

How many thesis were submitted in 2019? How many thesis were submitted in 2019 in the IRPS department?

Question 7

Rank the departments by number of theses written.

Question 8

The variable `thesis_language` contains other languages than English (en) and French (fr). Change all values that are not English or French to others.

Question 9

Economics is known for being the most globalized discipline. Which department has the highest share of thesis written in English?

Question 10

Subset the dataset for theses submitted between 1991 and 2020. Then, create a new variable called `thesis_decade`, which takes the following values:

- "90s" if submitted between 1991 and 2000;
- "00s" if submitted between 2001 and 2010;
- "10s" if submitted between 2011 and 2020.

(Tip: you can use `ifelse()`, but take a look at `dplyr::case_when()`.)

Finally, print a table with the number of theses in each department per decade:

Table 4: Theses per department-decade

thesis_department	1990s	2000s	2010s
department_1	n	n	n
department_2	n	n	n
. . .	. . .	. . .	. . .