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 PhD students over time. In this homework, we will work data to investigate how PhD theses at the Institute relate to gender and sustainability topics.

Question 1

The `topic.xlsx` file contains a variable that counts how many words are related to gender(`gen_degree`) and sustainability (`sus_degree`) in a theses abstract, but it is messy. Import the dataset and clean it so it contains only the following variables:

Name	Description
`thesis_ID`	ID of the Ph.D. thesis.
`sus_degree`	Number of words about sustainability in abstract.
`gen_degree`	Number of words about gender in abstract.

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 only 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 (*attention this is not the same dataset as last week*). Merge this dataset to the topic and departament datasets (from question 1 and question 2 above). The new "phd_theses" dataset should contain the following variables:

Name	Description
`thesis_title`	Title of the Ph.D. thesis.
`thesis_ID`	ID of the Ph.D. thesis.
`thesis_year`	The year in which a thesis was submitted.

Name	Description
`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 merged dataset.

a) How many departments are there?

b) In how many languages were theses written?

c) When was the first thesis written?

Question 5

a) How many thesis were submitted in 2019?

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

Question 6

Rank the departments by number of theses written.

Question 7

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

Question 8

Which department has the highest share of thesis written in English?

Question 9

a) Filter 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()`.)

b) Create 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
...	...	...	...

(Tip: pipe it from your code in 9a)

Question 10

Drop all observations with missing values for`sus_degree` and `gen_degree`. Then, calculate the yearly mean of `sus_degree` and `gen_degree` for years between 2017 to 2022. Display the results in a table like the following:

Table 5: Sustainaibility and Gender at IHEID Theses

year	gen_degree	sus_degree
2017	mean in 2017	mean in 2017
2018	mean in 2018	mean in 2018
2019	...	...
2020	...	...
2021	...	...
2022	...	...