Homework: cleaning and wrangling data

October 2022

Courses at the Graduate Institute

The Graduate Institute offers courses in spring and autumn. The datasets `autumn_21.csv` and `spring_22.csv`contain information on all courses offered by the Graduate Institute in the academic year 2021-2022.

Question 1

Open the `spring_22.csv` data and create a variable called "Department" with the department acronym for each course. (Tip: you have all the information necessary in the various dummy variables for department, you just need to pivot long the data and remove the NAs.)

Question 2

Open the `autumn_21.csv` data and create a variable called "Department" with the department acronym for each course. (Tip: here you will have to separate the acronyms for departments from the course code)

Question 3

Join the two datasets into one dataset called "academic_year". (Tip: remember to rename variables consistently across before joining data for best results).

The "acadmic_year" dataset should contain the following variables:

Name	Description
`title_course`	Title of the course
`department`	Department that offers the course (MINT, EI, RISP, HPI, DI...)
`language`	Language in which the course is instructed (French or English)
`ECTS`	How many ECTS you can get for the course.
`semester`	Takes the categories autumn or spring.
`type`	Type of course(compulsory, elective, or workshop)
`topic` A	broad category that summarizes the topic

Question 4

Do you have any duplicated rows in the "academic_year" dataset? If so, remove them.

Question 5

In the academic year, how many courses were offered in French at IHEID?

Question 6

In the academic year, how many courses were offered the in autumn semester and in English at IHEID?

Question 7

Rank the departments by the number of courses offered in each semester.

Question 8

Which department offers a higher share of courses in the spring semester? (Tip: after filtering and grouping, you need to dived the courses by department by the total number of courses in the psring semester).

Question 9

List the three favorite topics overall.

Question 10

List the three favorite topics of each department. (Tip: group and slice).

Question 11

One of the categories of `type` is "workshop". Workshops are normally about skills, but in the dataset, workshops are missing values for `topic`. Assign the category "skills" at the `topic` variable for all workshops.

Question 12

What is are the favorite topics for compulsory courses in all departments?

Question 13

Create a new dummy variable called "comp_type". The variable should take the value of 1 if a course is compulsory and about theory or methods; or take the value of 0 if a course is not compulsory or is compulsory but not about theory or methods.

Question 14

The `faculty_n.xlsx` dataset contains the number of faculty per department. In which department, faculty teaches more ECTs on average? Notice that faculty number for departments DE and IA are missing. (Tip: total ects per department divided by number of faculty).