Objects, class and data structures

Fundamentals of R Workshop - Homework 1

- 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.
- 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.

PhD theses at the Graduate Institute

The Institute has been home thousands of PhD students over time. In this homework, we will use a dataset on these PhD theses projects. Please download and load the "phd_thesis" dataset from the course Moodle page.

Question 1

a) How many columns does the dataset have?

b) How many rows does the dataset have?

Question 2

a) What is the class of the year variable in the dataset?

b) Please convert the year variable to numeric.

Question 3

a) How many PhD theses were written before the year 2000?

b) How many PhD theses were written in and after the year 2000?

Question 4

a) What is the mean for year?

b) What about the median?

c) Is the mean bigger than the median? Please create logical tests comparing the mean and median for year. Please also add a sentence explaining why do you think they differ.

Question 5

a) How many PhD theses were written in the PhD in Anthropology and Sociology program?

b) What is the mean year for PhD theses written in the PhD in Anthropology and Sociology program?

c) How many PhD theses were written in the PhD in International Relations/Political Science program?

d) What is the median year for PhD theses written in the PhD in International Relations/Political Science program?

Question 6

a) How many PhD theses were written in each program (hint: ´summary()´ and ´factor()´ could help)?

b) A lot of PhD theses do not have a program specified and are categorized as "No specialisation (1928-2001)". Could you please change this category within the program variable to "Not Available" (hint: brakets can help here or you can find help online).

Question 6

a) Suppose you wanted to know how many PhD theses were written about development, generally. Please select all the PhD theses that contain the word "development" in their title (hint: ´grep()´ could help you here).

b) Please create a new variable in the dataset called "about_development" with the value of 1 if the title of the PhD theses contains the word "development" in it or the value 0 if the PhD title does not contain the word "development" in it (hint: ´grepl()´ and ´ifelse()´ can help).

c) How many PhD theses that have the word "development" in their title were written before the year 2000? What about after the year 2000 (hint: ´subset()´ accepts multiple conditions)?

Question 7

a) For the PhD theses in the institute, candidates often have the choice of writing multiple essays or a manuscript. Suppose you are interested in their format, how many PhD theses contain the words "essay" or "essays" in their title and were written in the PhD in PhD in International Economics or the PhD in Development Economics programs (hint: you can use ´subset()´ with ´grepl()´)?

b) Please create a new variable called "essays_in_economics" with the value of 1 if a PhD theses contain the words "essay" or "essays" in their title and were written in the PhD in the PhD in International Economics or the PhD in Development Economics programs. Please add the value of 0 for everything else (hint: ´grepl()´ and ´ifelse()´ can help).

Questions 8

Please create a new dataset aggregating all PhD theses' titles by their program. The new dataset should have two columns, one called "program" with the PhD theses program and the other called "titles" with all the titles pasted together. The new dataset should have 12 rows, one for each program. Please use ´aggregate()´ and ´paste()´ to accomplish this. (hint: use the practical script for reference or find help online).