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 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.

PhD theses at the Graduate Institute

The Institute has been the home for hundreds of PhD students over time. In this homework, we will data about PhD theses at the Institute. Please download and load the "phd_theses" from the course Moodle page. It looks like this:

Name	Description
`title`	Title of the Ph.D. thesis.
`year`	The year in which a thesis was submitted.
`author`	The author of the thesis.
`department`	Department in which the thesis was written.
`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.
`complete_series`	Takes the value TRUE, if observation belongs to a year when all current departments already existed.

Question 1

a) What data structure is "phd_theses"?

b) Create a new vector called title that contains the title column from the phd_theses dataset.

c) How many observations does the dataset have?

d) How many variables 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 in 1930?

b) Did World War II (1939-1945) affect the number of theses written? Calculate the difference between the number of theses written in the seven years before the war and the seven years during the war.

Question 4

a) What is the mean for the variable `sus_degree`?

b) What about the mean for the variable `gen_degree`?

c) Please create logical tests comparing the mean of both variables and interpret the results.

Question 5

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

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

Question 6

a) How many PhD theses were written in each program ? Which department has been the most productive?

b) Some programs are older than others, for example, the ANSO department just turned 10 years while the IRPS department has existed for at least thirty. Subset the dataset to contain only observations for which `complete_series` is true, and count again the number of theses per department.

c) A lot of PhD theses do not have a department specified and are categorized as "other". Could you please change this category within the program variable to "Not Available" (hint: brackets can help here and you can find help here).

Question 7

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 thesis contains the word "development" or the value 0 if the PhD title does not contain the word "development" (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)? Please also add a sentence interpreting these results.

Question 8

For the PhD theses at 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 International Economics or Development Economics programs (hint: you can use ´subset()´ with ´grepl()´)?

Bonus Question

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 and find help here).