Problem Set 2

Introduction to R | University of Oxford Sociology

Problem Set 2

This problem set contains exercises from Session 2 that were originally done in small-groups. To reinforce your understanding, please complete these exercises independently. Answer the following questions using `R` in a Quarto document.

Exercise 1: Work with real-world data

For this exercise, download the CenSoc-Numident Demo file (as .CSV) and the accompanying codebook (as PDF) from the Harvard Dataverse or the course website. The CenSoc-Numident is an individual-level dataset with information on mortality and sociodemographic characteristics.

1.1 Read in the dataset using `read_csv()` from the tidyverse package.

1.2 How many columns are in the dataset?

1.3 How many rows are in the dataset?

1.4 List the column names. What are a few research questions that could be addressed using this dataset.

Exercise 2: Data manipulation

2.1 Filter the `censoc` data frame to include only women (sex == 2). Use the `filter()` command.

2.2 Filter the dataset to only include people born between 1905 and 1920 using the `byear` variable.

2.3 Select the columns `histid`, `death_age`, `sex`, and `ownershp`.

2.4 Calculate the average age of death for women (hint: refer to question 1).

Exercise 3 - Data visualization

3.1 Make a histogram of the variable `death_age`. When are most people dying?

3.2 Make a histogram of the variable `byear`. When are most people born?

3.3 Recode the variable `sex` from numeric values (1, 2) to take character values ("men" and "women"). Note that 1 = men, 2 = women.

3.4 Calculate the mean of of death for both men and women using `group_by()` and `summarize()`. Use the `death_age` variable. Do men or women live longer in this sample?

3.5 Make a histogram of the variable `death_age` for both men and women. Use the `filter()` command.

3.6 Now try adding the following line to the histogram you made in question 3.1: `+ facet_wrap(~sex)`

Exercise 4 - mortality advantage of homeowners

Do homeowners in the United States live longer than renters in the United States?

4.1 Using the `censoc` data.frame, create a new data.frame `censoc_homeownership` that filters out any "missing" values for the `ownershp` variable (missing = 0). Use the `filter()` command.

4.2 In the `censoc_homeownership` data.frame, create a new variable `homeowner` using the `mutate()` command and the `case_when()` command. Assign this new variable `homeowner` a value of "own" if `ownershp == 1` and a value of "rent" if `ownershp == 2`. Note: we can check the values for this variable here.

4.3 Make a histogram on the age of death for "homeowner" and "renter" groups using `ggplot` using the `censoc_homeownership` data.frame. Use the `+ facet_wrap(~homeowner)` command.

4.4 Calculate the average age of death for "homeowner" and "renter" groups. Which group lives longer, on average? Use the `group_by()` and `summarize()` functions. What are some possible explanations for homeowners living longer than renters in the US?