

Problem Set 2

Introduction to R | University of Oxford Sociology

Casey Breen

Problem Set 2

Complete the following questions in R within 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 [Harvard Dataverse](#). The CenSoc-Numident is an individual-level data with information on individual-level 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 1: 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 `ownership`
- 3.4 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 (1, 2) to take values “men” and “women”
- 3.4 Calculate the mean of of death for both men and women using `group_by()` and `summarize()`. Do men or women live longer?
- 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 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 “not available values” for the `ownership` variable (values of `ownership` = 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 `ownership == 1` and a value of “rent” if `ownership == 2`. Note: we can check the values for this variable here: https://usa.ipums.org/usa-action/variables/OWNERSHP#codes_section
- 4.3 Make a histogram on the age of death for “homeowner” and “renter” groups using `ggplot`. Use the `facet_wrap(~homeowner)` — and make sure you’re using your `censoc_homeownership` data.frame
- 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` commands. What are some possible explanations for homeowners living longer than renters in the US?