Problem Set 1

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.

1a

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

1b

How many columns are in the dataset?

1c

How many rows are in the dataset?

1d

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

Exercise 1: Data manipulation

2a

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

2b

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

2c

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

2d

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

Exercise 3 - Data visualization

3a

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

3b

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

3c

Recode the variable `sex` from numeric (1, 2) to take values "men" and "women"

3d

Calculate the mean of of death for both men and women using `group_by()` and `summarize()`. Do men or women live longer?

3e

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

3f

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?

4a

Google "IPUMS ownershp variable" and look at what each numerical value means. Recode `ownershp` to create a character variable `homeowner` that takes value "homeowner" or "renter". Filter out cases where we don't know whether someone was a homeowner or not.

4b

Make a histogram on the age of death for "homeowner" and "renter" groups using `ggplot`

4c

Calculate the average age of death for "homeowner" and "renter" groups. Which group lives longer, on average? Does this analysis tell us anything about homeownership and longevity?