Week 5: Coding in R III – Data Importing and Wrangling

Group work (2–3 students); Submit GitHub repo URL; Use the link to download materials needed
https://github.com/xucamel/Quantitative_course_week5/tree/main/HW

Objectives

- Practice reading and saving different types of data in R
- Use dplyr functions to process and summarise data
- Become familiar with bootstrap method and parallel computing.

Problem 1 – Reading Different Data Types

Download three files
https://github.com/xucamel/Quantitative_course_week5/tree/main/HW/Data

- A .csv file (fish.csv)
- An .xlsx file (fish.xlsx)
- An .rds file (fish.rds)

Tasks

1. Import all three into R.

2. Print the first five rows of each dataset.

Deliverables: R code

Problem 2 – Saving Data

Take one dataset you imported in Problem 1.

Tasks

1. Save it in three formats (i.e., .csv, .xlsx and .rds) in an Output fold.

2. Compare file sizes using file.info() for the three files containing the same data.

3. Write a short note: which format is best for (a) sharing, (b) compact storage?

Deliverables:

- R code and files saved in an output/ folder
- A short paragraph (in your script or README) with your comparison

Problem 3 – Wrangling Pipeline with dplyr

Use the dataset fish.csv for an end-to-end wrangling workflow. Use the pipe operator %>% to link Tasks 1–3, and save the result into an object called fish_output.

Tasks

1. Filter & Select
  - Keep only Walleye, Yellow Perch, and Smallmouth Bass in Lake Erie and Michigan
  - Keep columns: Species, Lake, Year, Length_cm, Weight_g.

2. Create Variables
  - Add Length_mm = Length_cm * 10.
  - Create Length_group using bins: ≤200, 200–400, 400–600, >600 mm (hint: mutate() + cut(...)) and Count how many fish fall into each Length_group by species

3. Summarise
  - For each Species × Year, calculate mean weight, median weight and sample size.

4. (Optional) Quick plot
  - A plot to show the temporal change of mean weight for each species.

6. Export Results
  - Save the new data to an Output folder (choosing any format).

Deliverables:
• R code and output file

Problem 4 – Reading Multiple Files at Once

Use folder Multiple_files that contains multiple survey data files. Each file is named by the year when the data were collected (e.g., fish_2017.csv, fish_2018.csv).

Tasks

1. Read all files in the folder at once and combine them into one data frame.

Deliverables: R code

Problem 5 – Parallel Computing for Bootstrap

Modify Bootstrap_parallel_computing.R .
https://github.com/xucamel/Quantitative_course_week5/tree/main/HW/Code

Note that it uses the file fish_bootstrap_parallel_computing.csv in the Data folder.
https://github.com/xucamel/Quantitative_course_week5/tree/main/HW/Data

Hint: You need to change the number of resamples and the sample size, and compute mean weight instead of length.

Tasks

1. Simulate resampling 10,000 times in Lake Erie, each time sampling 200 fish per species, and calculate the mean weight for each species.

2. Run the simulation in both serial and parallel mode.

3. Compare the running times and notice how many cores your computer has and uses for this task.

Deliverables: R code