

Lab4: Hands-on exploratory data analysis

2022-09-21

What you know already

In Labs 1-3 you learned:

- About three data types of variables : factors, strings or characters, numeric.
- How to use R as a calculator.
- About four data structures in R: vectors, matrices, data.frames, and lists. (the only missing is: arrays)
- How to create vectors, matrices, and data.frames and index their elements.
- How to create a basic R script with code and comments.
- How to read into data saved in a a tabular format (e.g. csv)
- How to have a general look at the imported data and understand its structure
- How to extract parts of the data for further analysis about another type of script to write R code and text: R markdown
- How to handling missing data in some of the functions we've learned so far.
- The ideal ways to save data files for analysis in R.
- A soft introduction to the Tidyverse suit of packages, focusing on the `dplyr` and `ggplot` packages.

Outline for today

- Catch up on DataCamp assignments from last week (there were A LOT) [~ 45 minutes]
- Apply what you learned with a guided case study + reviewing it together [~ 1:30 hours]
- Walk through some difficulties folks are having with LA1 and time for questions [~45 minutes]

Learning Outcomes

- Apply routine data wrangling tasks with `dplyr` core functions: `filter`, `select`, `mutate`, `arrange`, `transmute`, `summarise` to real dataset
- Use the pipe `%>%` to put these operations together.
- Choose the appropriate types of plots for a given data type

Tasks

- Complete the Data Camp Activities set as due 09/21.
- Work on the hands-on in class project
- Work on or ask questions about LA1
- Complete the R quiz 4 (will be posted by tomorrow on Moodle).

Data Camp activities to focus on, ordered by relevance

1. Introduction to the Tidyverse: Data Wrangling
2. Introduction to the Tidyverse: Grouping & Summarising
3. Data manipulation with dplyr: Transforming Data with dplyr
4. Introduction to the Tidyverse: Data Visualization
5. Exploratory Data Analysis in R: Exploring Categorical Data
6. Exploratory Data Analysis in R: Exploring Numerical Data
7. Reporting with R markdown: getting started with R markdown

8. Introduction to Data Visualization with ggplot2: Introduction
9. Reporting with R markdown: adding analyses and visualizations
10. Introduction to Data Visualization with ggplot2: Aesthetics

Applying what you've learned

In this working directory, you will find a data set that we will explore today. The `readme.txt` file contains information about what the columns mean.

Exercise 1: Mass vs Metabolism (graphing)

The relationship between the body size of an organism and its metabolic rate is one of the most well studied and still most controversial areas of organismal physiology. We want to graph this relationship in the `Artiodactyla` using a subset of data from a large compilation of body size data (Savage et al. 2004). The data are in the `mass_metabolism.csv` file in the `input_files/` directory in your workspace.

Make the following manipulations and plots with appropriate axis labels, or answer the questions.

- this is an R markdown file, so for each exercise, write your code inside a code chunk. You can create a code chunk by clicking at the “+C” green rectangle just above this area of your interface
- use `ggplot2` and `dplyr`
- after you make a plot, save it with a descriptive name as a `.png` using `ggsave`
- comment your code thoroughly

1: Make sure `ggplot2` and `dplyr` are installed. If they are not, install them. Then load both packages
 2: Read in the dataset and have a first look at the data:
 2.a: what are the dimensions (# rows, # columns)
 2.b: what are the classes of each variable?
 2.c: obtain a summary of each variable (column)
 2.d: check whether the dataset has any missing data

3. Explore the variable `body_mass`
 3.a: what is the range of the values?
 3.b: what does its distribution look like? (make a plot)
4. Explore the variable `metabolic_rate`
 4.a: what is the range of the values?
 4.b: what does its distribution look like? (make a plot)
5. Make a plot of metabolic rate and body mass: which variable should be the response variable and which should be the explanatory variable?
6. Make a plot of body mass vs. metabolic rate, with log10 scaled axes (this stretches the axis, but keeps the numbers on the original scale), and the point size set to 3.
7. Make the same plot as (5), but with the different families indicated using color.
8. The same plot as (5), but with the different families each in their own subplot.

Exercise 2 (if we have time)

Dr. Granger is interested in studying the factors controlling the size and carbon storage of shrubs. She has conducted an experiment looking at the effect of three different treatments on shrub volume at four different locations. The data is in the file `shrub-volume-data.csv` in the `input_files/` directory in your workspace.

1. Read in the dataset and have a first look at the data:
 1.a: what are the dimensions (# rows, # columns)
 1.b: what are the classes of each variable?
 1.c: obtain a summary of each relevant variable (column)
 1.d: check whether the dataset has any missing data
2. Explore the variables `length`, `width`, `height`:
 2.a: what are the ranges of the values?
 2.b: what does their distribution look like? (make a plot)
3. Select the data from the `length` column and print it out (using `select`).

4. Select the data from the `site` and `experiment` columns and print it out (using `select`).
5. Add a new column named `area` containing the area of the shrub, which is the `length` times the `width` (using `mutate`).
6. Sort the data by length (using `arrange`).
7. Filter the data to include only plants with heights greater than 5 (using `filter`).
8. Filter the data to include only plants with heights greater than 4 and widths greater than 2 (using `,` or `&` to include two conditions).
9. Filter the data to include only plants from Experiment 1 or Experiment 3 (using `|` for “or”). 10. Filter the data to remove rows with null values in the height column (using `!is.na`) 11. Create a new data frame called `shrub_volumes` that includes all of the original data and a new column containing the volumes (`length * width * height`), and display it.
10. Save your new data frame as a csv file called `shrub_volumes.csv`

References

Savage, Van M., et al. “The predominance of quarter-power scaling in biology.” *Functional Ecology* 18.2 (2004): 257-282.

Data Carpentry for Biologists. <https://datacarpentry.org/semester-biology/exercises/Dplyr-shrub-volume-data-basics-R/>

Data Carpentry for Biologists. <https://datacarpentry.org/semester-biology/exercises/Graphing-mass-vs-metabolism-R/>