Lab5: More hands-on!

2022-09-28

What you know already

In Labs 1-4 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 and index them.
- How to create a basic R script with code and comments.
- How to use markdown to write reports.
- 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
- 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.
- How to apply routine data wrangling tasks with `dplyr` core functions: `filter`, `select`, `mutate`, `arrange`, `transmute`, `summarise` to real dataset
- How to use the pipe `%>%` to put these operations together.
- Choose the appropriate types of plots for a given data type

Outline for today

- Catch up on DataCamp assignments [you need to do this before moving further]
- Apply what you learned in two case studies from last week [if you haven't done that already] - this will allos you to practice `dplyr` and `ggplot2` skills
- Learn about sampling in R

Learning Outcomes

- Apply routine data wrangling tasks with `dplyr` core functions: `filter`, `select`, `mutate`, `arrange`, `transmute`, `summarise` to real datasets
- Use the pipe `%>%` to put these operations together.
- Choose the appropriate types of plots for a given data type
- Explore the importance of random sampling
- Practice sampling from vectors or from data.frames

Tasks

- Complete the Data Camp Activities you haven't completed yet from previous weeks [these will be under "Past due"] ~ 40 min
- Complete Exercise 1: Mass vs Metabolism (graphing) ~ 30 min
- Complete Exercise 2: Size and carbon storage in shrubs ~ 30 min
- Complete Data Camp activity "Sampling in R: Introduction to Sampling" ~ 30 min
- Complete Exercise 3: Take random samples
- Complete the R quiz 5 (will be posted on Moodle).

Previously assigned 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 (from last week)

for each answer, include a chunk of code. Remember, this is a markdown file!

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 throughly;

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: size and carbon storage in shrubs (from last week)

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.

_*for each answer, include a chunk of code. Rememeber, this is a markdown file!_*

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 cvs file called `shrub_volumes.csv`

Exercise 3: Take random samples

For this you will use the dataset mentioned in Chapter 4 of your textbook and in lectures. The data is in the file `human_genes.csv` in the `input_files/` directory in your workspace.

IMPORTANT! First, complete this data camp activity about sampling (which will also make you use `dplyr` and `ggplot2`!): "Sampling in R: Introduction to Sampling". Then come back here and do this exercise.

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. Take a random sample of 10 genes (n=10) from this dataset. You can use the function `sample`. The code chunk below has the steps you need to complete to achieve this.

```
#1.first, read a bit about the sample.int function with ?sample.int

#2.now create a vector with numbers from 1 to the number of rows
#in the data.frame. Save that into an object.


#3.now use sample.int to sample 5 numbers randomly from this sequence of
#numbers. Save them into an object
```

```
#4.now use the vector with 5 numbers to index your data.frame (or tibble)
#by using the vector as the rows to index


#5.this retrieves the entire row corresponding to those numbers, but you can extract just the column si

#6.calculate the mean for those five "sizes"

#7.When you're done, repeat steps 3-6 (saving each step in a different object)

#8. Now compare the genes you sampled in each sampling event.
#Are they the same? How about the means?
#How do the means compare to the "real mean" of all human protein-coding genes?

#9. You can try a different way of sampling with dplyr's sample() function.
#Try it!


#10. Another way to do this is with dplyr's sample_n() function.
#Try it!
```

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/