Getting started with R – Graded Exercise

Submission deadline: 10th February

Exercise information and instructions for submission

To pass the course and receive the 0.5 ECT, you must submit an exercise that will be graded with pass or fail. After you have completed the exercise, you need to submit the following materials to us via email (carina.nigg2@unibe.ch & judith.bouman@unibe.ch):

- The R script
- The output (copy it into a word document)
- The data

Alternative: You can do this exercise also in an r Quarto or r Markdown file. These are files that, in contrast to a normal R script, combine code and output and allow annotation in an easy way. Due to time constraints, we have not covered this in this class, but there are online tutorials available on how to work with these types of documents, for instance: https://quarto.org/docs/get-started/hello/rstudio.html

If you decide to work with Markdown / Quarto, please submit the rendered html-file (or pdf) to us – in this case, we do not need the dataset.

To pass the exam,

- you must have completed each of the exercises below,
- you must have submitted the script with the solution code for the exercises, the output, and the data to us (no need for data if you work with R Quarto or R Markdown and share a rendered html-file),
- when we run the code on your dataset, it must work. More than 3 errors will result in fail.

Graded exercise

Complete the following exercises with a dataset of your choice. Clearly annotate all your exercises and show us where you do what, for example, by adding headers (# Exercise 1) and annotating functions (e.g., # calculating the median). Most of the functions we have covered in the class; however, R is an environment of constant learning and there might be one or two functions for which you cannot find code in the exercises we did. This means you have to find your own solution on how to do this.

Exercise 1 Set up folder structure and load packages

1. Set up the project root directory. Within the root directory, create the following folder structure using the R code.
   - 01_data: For your dataset(s).
   - 02_code: For your R script or markdown/Quarto file.
   - 03_output: For the findings of your analysis: figures, tables, or a summary of your findings.
2. Check the project directory.
3. Install (if necessary) and load all required packages you will work with.

Exercise 2: Import your dataset and check structure

1. Import your dataset into R using an appropriate function.
2. Provide an overview of columns and rows in your dataset.
3. Display the first few rows of the dataset.

Exercise 3: Clean and transform your data

Use tidyverse for all the following exercises.

1. From your imported dataset, select only the variables relevant to your analysis. These should include at least two categorical and two numeric variables.
2. Check the class of the variables you will be working with for your analysis. There must not be any strings in your key variables. If required, adjust the class to the correct one.
3. Arrange your dataset in descending order based on one of the numeric variables.
4. Create a new dataframe with a new name that consists of a random selection of 5 rows of your full dataset. (hint: this is not included in the class exercises)
5. Rename at least two variables. Use short names and lower case.
6. Create at least the following two new variables:
   I. A ratio or percentage variable using existing columns.
   II. A categorical variable from a numeric one. The categorical variable should have at least three categories. Make sure that there are observations for each category (so no category with 0 cases). Make sure to assign appropriate value labels for the categories.
7. Move the new categorical you created (exercise 3.6) after the numeric one which it is based upon.

8. Check levels for one of your categorical variables. Change the reference category (which is the first one) to a different one.
9. Check for duplicates in your dataset. (hint: this is not included in the class exercises)

Exercise 4: Exploratory data analysis

1. Provide appropriate summary statistics for your key variables (at least two numeric and two categorical ones).
2. Provide an overview of missing data for your key variables if not done in the previous exercise (4.1).
3. Calculate group-wise summaries for one numeric variable, this should include mean, median, standard deviation, min, and max, grouped by a categorical variable.
4. Create a frequency and proportion table for two categorical variables.

Exercise 5: Data visualization with Base R

Use **Base R** for the following visualizations.

1. A histogram for a numeric variable.
2. A boxplot for a numeric variable, grouped by a categorical variable. Provide a title for your plot, label your x-axis, and specify the colors for each stratum.
3. A bar chart for a categorical variable.
4. A scatterplot for two numeric variables.