Exercise Sheet #3

Fortgeschrittene Statistische Software für NF

Lisa Bondo Andersen & Anna Steinberg

2025-05-23

General Remarks

- You can submit your solutions in teams of up to 3 students.
- Include all your team-member's names and student numbers (Matrikelnummern) in the `authors` field.
- Please use the exercise template document to work on and submit your results.
- Use a level 2 heading for each new exercise and answer each subtask next to its bullet points or use a new level 3 heading if you want.
- Always render the R code for your solutions (`echo=TRUE`) and make sure to include the resulting data in your rendered document.
  - Make sure to not print more than 10 rows of data (unless specifically instructed to).
- Always submit both the rendered document(s) as well as your source Rmarkdown or Quarto document. Submit the files separately on moodle, **not** as a zip archive.
- Submission format is HTML. Other formats will lead to a deduction of points.

Exercise 1: Initializing git (4 Points)

For this whole exercise sheet we will be tracking all our changes to it in git.

a) Start by initializing a new R project with git support, called `2025-exeRcise-sheet-3`. If you forgot how to do this, you can follow this guide.
b) Commit the files generated by Rstudio.
c) For all of the following tasks in this exercise sheet we ask you to always commit your changes after finishing each subtask e.g. create a commit after task *1d*, *1e* etc.

   Note: This applies only to answers that have text or code as their answer. If you complete tasks in a different order or forget to commit one, this is no problem. If you change your answers you can just create multiple commits to track the changes.

d) Name 2 strengths and 2 weaknesses of git. (Don't forget to create a commit after this answer, see *1c*)
e) Knit this exercise sheet. Some new files will automatically be generated when knitting the sheet e.g. the HTML page. Ignore these files, as we only want to track the source files themselves. You can, but don't need to create a `.gitignore` file. Just do not commit these files manually.

Exercise 2: Putting your Repository on GitHub (3 Points)

For this task you will upload your solution to GitHub.

a) Create a new repository on GitHub in your account named `exeRcise-sheet-3`. Make sure you create a **public repository** so we are able to see it for grading. Add the link to the repository below:
b) Push your code to this new repository by copying and executing the snippet on github listed under `...or push an existing repository from the command line`.
c) Regularly push your latest changes to GitHub again and especially do so when you are finished with this sheet.

Exercise 3: Pixar Films (4 Points)

Download the `pixar_films` and `public_response` datasets from the GitHub repository and track them in git.

Link: https://github.com/rfordatascience/tidytuesday/tree/main/data/2025/2025-03-11

For small datasets like these adding them to git is not a problem.

a) Load the `pixar_films` dataset into R. Clean the dataset by removing films without a title. Inspect the variable `film_rating`. What are the possible values and what do they mean? Create a factor variable for the film rating. Why is this appropriate?

b) Inspect the film titles manually. Which films form a film series? A film series can be identified by a common word in the titles of the films, often in conjunction with a number in the title, e.g. "Despicable Me" and "Despicable Me 2". Create a dataframe which displays a list of the different series with the titles of the films and how many films belong to the series. Output the dataframe.

c) Load the `public_response` dataframe into R. Convert the `cinema_score` variable into a factor while ensuring the factor levels are defined in ascending order, from the lowest to the highest score. Combine `public_response` with the `pixar_films` dataset using an appropriate merge variable.

d) Choose one of the variables representing the public response and create a bar plot for the films belonging to a series. Here are the details of the plot:

- The film series are represented on the x-axis.
- Your chosen public response variable is displayed on the y-axis.
- Each film in the series is represented as a separate bar. Bars are grouped by film under their respective series on the x-axis. Order the bars within a series according to the release date of the films.
- A title and axis labels for context.

What do you notice when comparing the scores of the films in a series? Do you see any patterns?

Exercise 4: Open Analysis (4 points)

This exercise is a bit more open-ended. You can choose any dataset from Our World in Data and analyze it, while determining the research question yourself.

a) Go to https://github.com/owid/owid-datasets/tree/master/datasets and choose a dataset that interests you. You can have a look at https://ourworldindata.org/ to gather some inspiration.
b) Download the dataset and track it in git.
c) Put the name / title of the dataset and a link to it below.

- Dataset Name: ...
- Link: https://github.com/owid/owid-datasets/...

d) Come up with a (research) question you want to answer with the data and briefly explain why you believe this is an interesting question within one sentence. It should be a question that can be answered with the dataset and using R.
e) Use R to answer your chosen question. Please limit your analysis to the functions and techniques we have covered so far in the course. You are **not expected** to use advanced statistical models or external packages which haven't been introduced.
f) Create a meaningful plot / figure with the dataset. Make sure to provide a figure caption (via the chunk options / Rmarkdown) and correctly label the figure.

Final Note

Make sure to push all your commits and changes to GitHub before submitting the exercise sheet.