D1 — Handling and Visualizing Data
Preparation for Final Assessment of Learning Outcome, AS2024

The final assessment of learning outcome will be an individual project work carried out during the last central teaching. This document gives you an idea of how the project will look like.

In the sections below you find questions and tasks (indicated by the letter **T** followed by an index) that you should carry out. For the actual exam project we will indicate the number of points you can get for each task.

You will carry a data set through the various steps treated in D1 (inspection, loading into R, cleaning, visualization with ggplot, design considerations for your plot, numeric measures), thereby applying the knowledge gained in D1. For the preparation we provide the data-set *mock_project_data.csv* on Moodle. For the actual exam project, you will select and download a data-set from resources we provide.

During the actual exam project, you will have to submit a data file, your R code, and a visualization via Moodle. The R code should be structured according to the tasks T1.1—T7.3, in the correct order. For each task provide an R-comment section (i.e. with #) starting with the task number, followed by a brief answer to the corresponding question. After the comment, provide the corresponding R code, if required by the task.

Here is an example where R code is required:

```
# T5.1
#
# My basic plot is a barplot. Each bar corresponds to a dose and shows the number of sick days.
# Note: stat=”identity” is required to calculate the sum of the y variable, grouped by the x variable

p = ggplot(data=df, aes(x=dose, y=days)) + geom_bar(stat="identity")
```

Note that some tasks (like T2.1, or T5.1—T5.3) will only require an answer without R code. In this case, you will provide an R-comment with the task number, followed by your answer. For example

```
# T2.1
#
# My CSV file was not a pure table; it contains a caption. I have opened the file with my
# text editor (OS X’s TextEdit) and removed the caption.
```

Note that during the exam project not every task will necessarily apply to your chosen data-set. For example, some data sets are already tidy, or do not contain missing data, or they do not require the conversion of types.

1. Find a data set

T1.1 Download *mock_project_data.csv* from Moodle.

2. Inspect the data set with a text editor and load it into R

Inspect your CSV file with a text editor (e.g. TextEdit on OS X, Notepad on Windows) before loading it into R. We do not recommend inspecting the file with Excel (or some other spread sheet software). Do not edit the file with Excel and save it as CSV; this may mess up the data, as spread sheet software sometimes converts the format of certain values to a wrong type.

T2.1 Some data sets are not a pure CSV, because they contain a title and/or a caption. Remove them using your text editor. If your data set is a pure CSV, skip this step.

T2.2 Load your CSV file into R. You can use the function *read.delim*. Pay attention to supply correct arguments concerning the separator between columns and decimal point, and for missing data.

3. Clean the data set and inspect it

This step *may* involve tidying the dataset (using the *tidyverse* package), handling missing data, converting variables to their proper types, or removing inconsistencies.

T3.1 If your data set is not tidy, tidy it.
T3.2 If one or more variables have an incorrect type, convert to the correct type.
T3.3 What are the variables? What is the type of each variable?
T3.4 How many observations are in the dataset?
T3.5 What is the pattern of missing data? Use the function *md.pattern* of the package *mice* and provide an interpretation of the output.
T3.6 If your data set has missing data, treat it. Possible approaches are deletion and imputation.
T3.7 This is an open task. It is for any pre-processing that you do that does not fall into any of the above tasks. An *example* is if you must compute cumulative of counts for your intended plot. Describe what you did and provide the corresponding R code.

4. Visualize your data with ggplot

You *must use ggplot*. For each task in this section describe in your report *briefly* what you did (with one or two sentences for each task). Here is an example for T4.3: *"I made small multiples, splitting the data according to the categorical variable Country."*

T4.1 Construct a basic plot, i.e., a plot with one layer only, and without faceting.
T4.2 Add *an additional layer* or apply the technique of *small multiples* to your basic plot from T4.1

For the following tasks use the plot from T4.2 and carry out the following tasks:

T4.3 Label the aesthetics (the axes and scales) with a meaningful label.
T4.4 Add a title to your plot.
T4.5 Add a short caption to the plot.
T4.6 This is an open task. Add modifications to your plot. You can, *for example*, introduce further style changes to title and/or axis label(s) and/or caption, making them bold, or cursive, or in color, or change the font-size. Or you can change the place in the plot where they are shown, put margins around the plot, apply a theme, or set the aspect ratio. Describe what you did and provide the corresponding R code.
T4.7 Save your plot in PDF format using an R command.

Code of conduct for the exam project

- You are *not* allowed to use R code that has been produced by others for the analysis of COVID-19 data.
- You are allowed to take inspiration from example R code found on the Internet, but you will have to personalize and adapt it to your own analysis.
- The R code that you submit must be your own work.
- You are not allowed to support other students in producing their answers and R code.
- You are allowed to use chat bots like chatGPT. Please provide the query you used as a comment in the R file under the corresponding task.
- Note that your proposed solutions must be thought through. Do not just copy-paste code that is somehow related to the task, in the hope to get points. In fact, we may deduct points for semi-sensical copy-pasted code.