Notes for Getting Data Analysis Support

Data File

GIGO also affects data analysis!

1. **Variables (e.g., sex, age) in columns, participants in rows!**

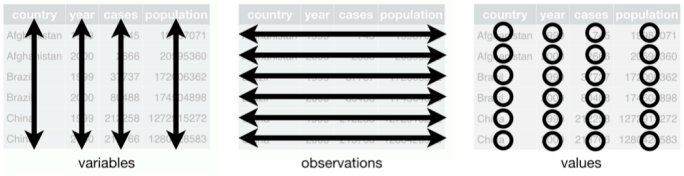

Image: Grolemund & Wickham (2017)

2. **The first line (and only the first) contains the variable names!**
   Short names without spaces (if needed, use _ for spaces). As short as possible, as long as necessary to understand it. Values start in the second row.
3. **Every participant must have values only in one row!**
   If the participants can take part multiple times, ask how you format the data.
4. **Use , decimal sign if necessary (do not mix with .)**
5. **Use values consistently!**
   Not "male" and "man" in the same column.
6. **Fileformat ideally .csv (; as divider, "" for text)**
   Excel can expor the current worksheet as .csv. Use Unicode (UTF-8). If Excel exports additional empty rows, delete these empty rows (Notepad++, TextWrangler/BBEdit can be used). SPSS (.sav) can export as .csv as well.
7. **Data file must only contain the variable names in the first row and the values of the participants in the others.**
   No means, sums, or other comments in the file!

Codebook

The codebook is necessary to understand what the variables and values mean. It consists of:

1. **Consecutive Number (column in data file)**
2. **Variable names in the data file (see above)**
3. **Complete variable name**
4. **Range of values (incl. unit of measurement for continuous variable)** *or* **possible values (for discrete variables; if values are coded with number, the number incl. its meaning)**
   If vales are used for "n/a", "missing values", etc. (e.g., 999) mention them as well!

Example

Whole data file: missing values = 999

4.	age	Age	0-120 years
5.	sex	Sex	1 = male, 2 = female
6.-14.	ati01-ati09	ATI-Questionnaire (Franke, Attig, Wessel, 2018) Values: 1-6 (1 completely disagree, 2 largely disagree, 3 slightly disagree, 4 slightly agree, 5 largely agree, 6 completely agree)	

Questions to the Data

Write down the questions you want to have answered by the data (mention the variable names as well).
Differentiate between

1. **Questions regarding relationships**
   E.g. "Do react people slower (answer_time) the older (age) they are?"
2. **Questions regarding group differences**
   E.g., "Is the red website (grade_red) evaluated better than the blue website (grade_red)?" (if participants evaluated both websites)
   E.g., "Are the visitors of the red website more satisfied than the visitors of the blue website?" (satisfaction: satisfaction; group variable: website_visited)

Evaluation/Study/etc. Procedure

What was done with the participants/what did the participants do and when did they do it — sketch timeline.

Person is randomly assigned to one condition or participant enters the usability lab.

What was done with the participants? What was collected?

Person exists the website or the usability lab.

The data analysis can only be as good as how well the study was conducted.

Please discuss how you are going to collect the data beforehand.

Otherwise you waste your time, my time, and (even worse) your participants' time.