**Template for data cleaning checklist**

|  |  |
| --- | --- |
| Author: |  |
| Date: |  |
| Version number: |  |

|  |  |
| --- | --- |
| **1. Importing the data** | **What type of import format will be used?**  **What cleaning steps are done before importing into R?**  **Where do the raw data files come from?** |
| Steps to perform: |  |
| Packages used: |  |
| **2. Cleaning column names** | **Clean column header.**  **What is the style for variable names?**  **Which regular expression denote different time points or raters?** |
| Steps to perform: |  |
| Packages used: |  |
| **3. Dealing with structural problems** | **Is melting/casting necessary?**  **Define the tidy structure of the dataset (long or wide format).**  **Can you keep data in one master file or do you need several master files?** |
| Steps to perform: |  |
| Packages used: |  |
| **4. Column creation or transformations** | **Which columns need to be created?**  **Which transformations need to run?**  **Always base calculations on other variables and code them in R (never do them outside the project or you will have trouble replicating steps).**  **Remember to never overwrite existing columns.** |
| Steps to perform: |  |
| Packages used: |  |
| **5. Selecting or re-ordering columns** | **Create a logical order of variables in your datafile.**  **Consider building an order that involves id variables – independent variables – effect modifiers (moderating variables) – dependent variables** |
| Steps to perform: |  |
| Packages used: |  |
| **6. De-duplication** | **If you suspect duplicated rows of observations, run unique() and the like.** |
| Steps to perform: |  |
| Packages used: |  |
| **7. Re-coding values** | **Re-coding involves: Splitting up numeric variables into categories or**  **Defining alternative categories in categorical variables (factor variables).**  **Never overwrite existing variables. Create re-coded variables as new variables.** |
| Steps to perform: |  |
| Packages used: |  |
| **8. Setting up factor variables** | **Set up factor variables which you haven’t set up already under (4) or (7).**  **Define factor levels of cut variables (numeric variables put into categories) from step (7).** |
| Steps to perform: |  |
| Packages used: |  |
| **9. Setting up date variables** | **Set up date variables using lubridate.**  **Create time difference variables if necessary.** |
| Steps to perform: |  |
| Packages used: |  |
| **10. Row-wise calculations** | **Perform row-wise calculations (i.e., rowwise sums or means, also rowwise calculations for grouped data).** |
| Steps to perform: |  |
| Packages used: |  |
| **11. Final sort and arrange** | **Sort your data according to id variables or other values in variables to have a clean order of observations in your file.**  **You might want to do another arrange after the last calculations.** |
| Steps to perform: |  |
| Packages used: |  |
| **12. Export the masterfile** | **Export the masterfile into the “data“ folder.**  **Always use .rds as masterfile format.** |
| Steps to perform: |  |
| Packages used: |  |
| **13. Update the codebook** | **Update the codebook in the “munge” folder.** |
| Steps to perform: |  |
| Packages used: |  |

After you have planned your data cleaning, translate the structure of this document into the munge\_project\_01.R file in the munge folder of your project.