The Journal of Open Source Software

sjmisc: Data and Variable Transformation Functions

Daniel Lüdecke[1]

1 University Clinical Center Hamburg-Eppendorf

DOI: 10.21105/joss.00754

Software

- Review
- Repository
- Archive

Submitted: 18 May 2018
Published: 20 June 2018

Summary

Data preparation is a common task in research, which usually takes the most amount of time in the analytical process. There are typically two types of data transformation: arranging and reshaping data sets (like filtering observations or selecting variables, combining data sets etc.) and recoding and converting variables. Statistical software packages should provide convenient tools to fulfil these tasks.

For the *R Project for Statistical Computing*, packages have been released recently that are known to be part of the *tidyverse*. Some of those packages focus on the transformation of data sets. Packages with special focus on transformation of *variables*, which fit into the workflow and design-philosophy of the tidyverse, are missing.

`sjmisc` is a package for the statistical progamming language **R**, which tries to fill this gap. Basically, this package complements the `dplyr` package (Wickham et al. 2017) in that `sjmisc` takes over data transformation tasks on variables, like recoding, dichotomizing or grouping variables, setting and replacing missing values, etc.

The data transformation functions in this package all support *labelled data* (or labelled vectors), which is a common data structure in other statistical environments to store meta-information about variables, like variable names, value labels or multiple defined missing values. Working with labelled data is featured by packages like `haven` (Wickham and Miller 2018) or `sjlabelled` (Lüdecke 2018a).

The design of data transformation functions

The design of data transformation functions in this package follows, where appropriate, the tidyverse-approach, with the first argument of a function always being the data (either a data frame or vector), followed by variable names that should be processed by the function. If no variables are specified as argument, the function applies to the complete data that was indicated as first function argument. This design-philosophy makes it possible to combine functions from `sjmisc` and the "pipe-workflow", i.e. to create chains of function calls connected with `magrittrs` pipe-operator.

Conversion of Variable Types

There are also functions that convert variable types, e.g. from factors to numeric (or vice versa). These functions mimic R base functions, but also share the previously mentioned advantages of supporting labelled data and integrating seamlessly into the well-known pipe-workflow from tidyverse-packages.

The source code for `sjmisc` has been archived to Zenodo and linked with a DOI (see Lüdecke 2018b).

References

Lüdecke, Daniel. 2018a. "Sjlabelled: Labelled Data Utility Functions," May. Zenodo. https://doi.org/10.5281/zenodo.1249216.

———. 2018b. "Sjmisc: Data and Variable Transformation Functions," May. Zenodo. https://doi.org/10.5281/zenodo.1249192.

Wickham, Hadley, and Evan Miller. 2018. *Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files.* https://CRAN.R-project.org/package=haven.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Müller. 2017. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.