

datawizard: An R Package for Easy Data Wrangling and Transformations

Daniel Lüdecke¹, Dominique Makowski², Mattan S. Ben-Shachar³,
Brenton M. Wiernik⁴, Etienne Bacher⁵, and Indrajeet Patil⁶

¹ University Medical Center Hamburg-Eppendorf, Germany ² Nanyang Technological University, Singapore ³ Ben-Gurion University of the Negev, Israel ⁴ Facebook ⁵ Luxembourg Institute of Socio-Economic Research, Luxembourg ⁶ esqLABS GmbH

DOI:

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Submitted:

Published:

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Summary

The `{datawizard}` package in the R programming language ([R Core Team, 2021](#)) provides a

Statement of Need

The `{datawizard}` package makes basic data wrangling easier than with base R. Its workflow and syntax are designed to be similar to `{tidyverse}` (Wickham et al. (2019)), which is a widely used ecosystem of packages for data analysis, and, therefore, users familiar with this ecosystem can easily translate their knowledge. Naturally, one might wonder why recreate data wrangling functionality already present in `{tidyverse}`.

The `{easystats}` (Ben-Shachar et al. (2020), Lüdecke et al. (2020), Lüdecke, Ben-Shachar, et al. (2021), Lüdecke, Patil, et al. (2021), Lüdecke et al. (2019), Makowski et al. (2019), Makowski et al. (2020)) is an ecosystem of packages designed to make statistical analysis easier in R. Importantly, in order to be lightweight, it follows a “0-external-hard-dependency” policy. Thus, while building this ecosystem, a new data wrangling package that relies only on base R needed to be created. In effect, this package provides the data processing backend for this entire ecosystem. In addition to its usefulness to the `{easystats}` ecosystem, it also provides *an* option for R users and package developers if they wish to keep their (recursive) dependency weight to a minimum (for other options, see Dowle & Srinivasan (2021), Eastwood (2021), etc.).

In addition to providing functions to clean messy data, `{datawizard}` also provides helpers for the other important step of data analysis: transforming the cleaned data further for setting up statistical models. For example, one may need to standardize certain variables, normalize range of some variables, adjust the data for effect of some variables, etc.

Lastly, `{datawizard}` also provides a toolbox to create a detailed profile of data properties.

Features

Data wrangling

Table 1: The table below lists a few key functions offered by *datawizard* for data wrangling. To see the full list, see the package website: <https://easystats.github.io/datawizard/>

Function	Operation
<code>data_filter()</code>	to select only certain <i>observations</i>
<code>data_select()</code>	to select only a few <i>attributes</i>
<code>data_extract()</code>	to extract a single <i>attribute</i>
<code>data_rename()</code>	to rename attributes
<code>reshape_longer()</code>	to convert data from wide to long
<code>reshape_wider()</code>	to convert data from long to wide
<code>data_join()</code>	to join two data frames
...	...

Data transformations

Table 2: The table below lists a few key functions offered by *datawizard* for data transformations. To see the full list, see the package website: <https://easystats.github.io/datawizard/>

Function	Operation
<code>standardize()</code>	to center and scale data
<code>normalize()</code>	to scale variables to 0-1 range
<code>adjust()</code>	to adjust data for effect of other variables
<code>data_shift()</code>	to shift numeric value range
<code>ranktransform()</code>	to convert numeric values to integer ranks
...	...

Data properties

The workhorse function to get a comprehensive summary of data properties is `describe_distribution()`, which combines a set of indices (e.g., measures of centrality, dispersion, range, skewness, kurtosis, etc.) computed by other functions in `{datawizard}`.

```
describe_distribution(mtcars$wt)
#> Mean | SD | IQR | Range | Skewness | Kurtosis | n | n_Missing
#> -----
#> 3.22 | 0.98 | 1.19 | [1.51, 5.42] | 0.47 | 0.42 | 32 | 0
```

Licensing and Availability

datawizard is licensed under the GNU General Public License (v3.0), with all source code openly developed and stored at GitHub (<https://github.com/easystats/datawizard>), along with a corresponding issue tracker for bug reporting and feature enhancements. In the spirit of honest and open science, we encourage requests, tips for fixes, feature updates, as well as general questions and concerns via direct interaction with contributors and developers.

Acknowledgments

datawizard is part of the collaborative [easystats](#) ecosystem. Thus, we thank the [members of easystats](#) as well as the users.

References

- Ben-Shachar, M. S., Lüdtke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56), 2815. <https://doi.org/10.21105/joss.02815>
- Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of 'data.frame'*. <https://CRAN.R-project.org/package=data.table>
- Eastwood, N. (2021). *Poorman: A poor man's dependency free recreation of 'dplyr'*. <https://CRAN.R-project.org/package=poorman>
- Lüdtke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, 5(53), 2445. <https://doi.org/10.21105/joss.02445>
- Lüdtke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
- Lüdtke, D., Patil, I., Ben-Shachar, M. S., Wiernik, B. M., Waggoner, P., & Makowski, D. (2021). see: An R package for visualizing statistical models. *Journal of Open Source Software*, 6(64), 3393. <https://doi.org/10.21105/joss.03393>
- Lüdtke, D., Waggoner, P., & Makowski, D. (2019). insight: A unified interface to access information from model objects in R. *Journal of Open Source Software*, 4(38), 1412. <https://doi.org/10.21105/joss.01412>
- Makowski, D., Ben-Shachar, M. S., & Lüdtke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, 4(40), 1541. <https://doi.org/10.21105/joss.01541>
- Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdtke, D. (2020). Methods and algorithms for correlation analysis in R. *Journal of Open Source Software*, 5(51), 2306. <https://doi.org/10.21105/joss.02306>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>