The Journal of Open Source Software

overviewR - Easily Explore Your Data in R

Cosima Meyer[1] **and Dennis Hammerschmidt**[1]

1 University of Mannheim, Germany*

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

The package overviewR (Meyer and Hammerschmidt 2022) helps users to get an overview of data with a particular emphasis on the extent that the distinct units of observation are covered for the entire time frame of the data set. With (large) data sets that have many different observations over a long period, it becomes increasingly difficult to identify each unique observation and its exact coverage in the data. In particular, if some observations are not included for the entire time span of the data – either because they entered later, dropped out earlier, or have gaps in between – it can become difficult to spot potential problems in the data's time and scope. As Staniak and Biecek (2019) argue, these tools are increasingly in demand.

The main advantages of overviewR are threefold: 1) it allows users to quickly understand the data and the distribution of the observations over time, 2) it offers several helper functions that facilitate complex variable engineering tasks such as the merging of multiple data frames on different time formats or the generation of specific time periods, and 3) it allows to show this information both in a visual and tabular form that is convenient for academic publications.

Statement of need

overviewR has already been used in the data wrangling process as well as to generate an illustrative overview of the sample in various academic publications and research (Hammerschmidt 2021; Hammerschmidt, Meyer, and Pintsch 2021; Meyer and Bolin 2022; Meyer 2021). While not limited to these applications, we see its main advantage for researchers and practitioners working with time-series cross-sectional data and its power to illustratively visualize key information of the data. When comparing with other exploratory data analysis packages available (for instance smartEDA (Putatunda et al. 2019), dlookR (Ryu 2022), or summarytools (Comtois 2022)), we do not see the key functions of overviewR covered.

Key functions

overviewR can be used by everyone who works with data that have time-and-scope characteristics. To get a quick overview of which units (i.e. countries, companies, test persons, etc.) are present or missing during a given time span (i.e. years, months, days, minutes, etc.), overviewR provides an easy and intuitive insight into the set-up of your data. But

*Both authors are currently working as data scientists in the industry.

The Journal of Open Source Software

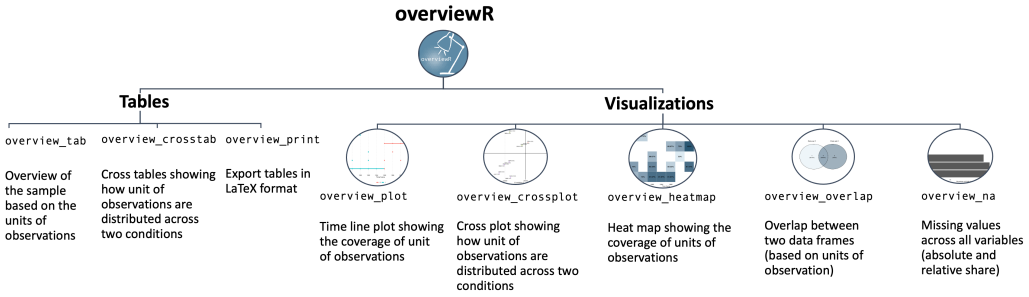

Figure 1: Comparing the functions of overviewR

overviewR goes beyond this: It also allows users to investigate logical clusters of time-unit observations by using cross tables (or cross plots for visualization). It further allows the user to investigate the overlap between two data sets as well as the distribution of missing values across the data. overviewR relies on `ggplot2` (Wickham, Chang, et al. 2022) and `ggvenn` (Linlin 2021) for visualization. It comes with example data containing a simple time-series cross-sectional data set. All functions work with `data.frame` objects while `overview_tab` and `overview_na` can also handle `data.table` objects to increase the performance.

As Figure 1 shows, the key functions of overviewR can be divided into its type of presentation (tabular or visual). They may, however, also be categorized along their functionalities: 1) a general overview of the units of observation (`overview_tab`, `overview_plot`, `overview_heatmap`), 2) a logical cluster within the data (`overview_crosstab`, `overview_crossplot`), 3) key information about the data sets (`overview_overlap`, `overview_na`), and 4) export functions (`overview_print`).

Visualization of a workflow

To visualize a typical workflow, we rely on the internal data provided by overviewR.

```
data(toydata)
head(toydata)
```

```
#   ccode year month      gdp population day
# 1   RWA 1990   Jan 24180.77  14969.988   1
# 2   RWA 1990   Jan 23650.53  11791.464   2
# 3   RWA 1990   Jan 21860.14  30047.979   3
# 4   RWA 1990   Jan 20801.06  19853.556   4
# 5   RWA 1990   Jan 18702.84   5148.118   5
# 6   RWA 1990   Jan 30272.37  48625.140   6
```

There are 264 observations for 5 countries (Angola, Benin, France, Rwanda, and UK) stored in the`ccode`variable, over a time period between 1990 to 1999 (`year`) with additional information for the month (`month`). Additionally, two artificially generated fake variables for the gross domestic product (GDP, `gdp`) and population size (`population`) are included to illustrate conditions.

We can now use `countrycode` (Arel-Bundock et al. 2022) as well as `dplyr` (Wickham, François, et al. 2022) to transform the `ccode` into more meaningful information and plot a visualization using `overview_plot`:

```
toydata %>%
  # Transform the country code (ISO3 character code) into
  # a country name using the `countrycode` package
  dplyr::mutate(country =
                  countrycode::countrycode(ccode, "iso3c",
                                           "country.name")) %>%
  overview_plot(id = country, time = year)
```

The plot shows the sample distribution of all countries in the sample and illustrates at which year data is (consecutively) available and when years are missing (see Figure 2).

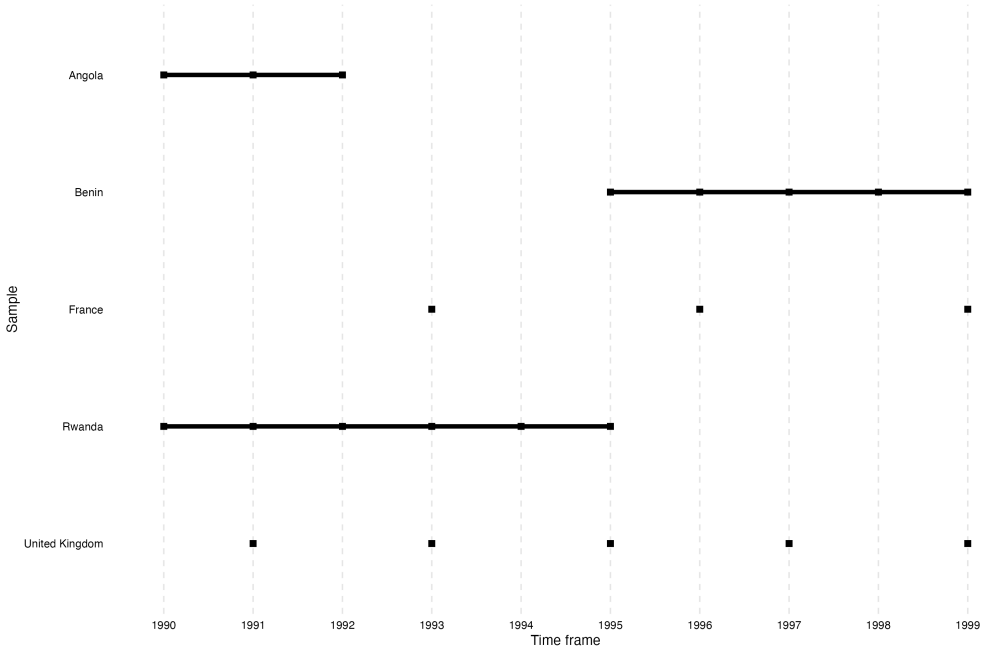

Figure 2: Visualization of 'overview_plot'

Going further, we can then also rely on `overview_heat` to identify the coverage of each month per country-year observation.

```
overview_heat(toydata_red,
              ccode,
              year,
              perc = TRUE,
              exp_total = 12)
```

Taking a twelve-month coverage as the baseline (with `exp_total`), we see in Figure 3 that only Benin (`BEN`) in 1997 achieves full coverage. With this function, researchers can also analyze the coverage of panel (survey) data.

If we wanted to compare meaningful clusters within the sample, we can rely on `overview_crossplot`. Similar to a tabular cross table (as presented in function `overview_crosstab`), this function visualizes the units of observations in the package across two conditions. The user can define and set the conditions as well as the thresholds as shown in the following example.

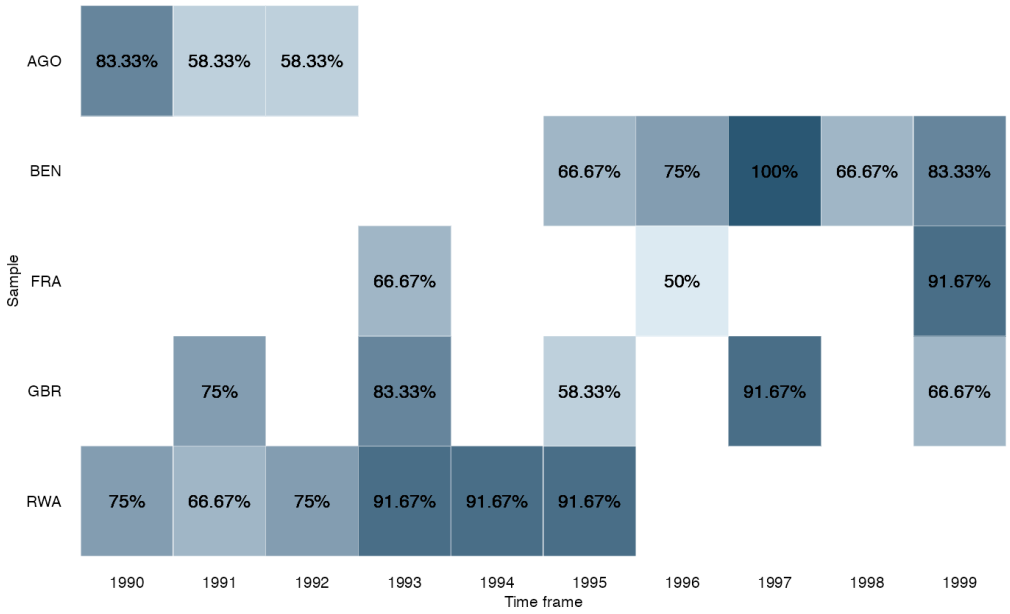

Figure 3: Visualization of 'overview_heat'

```
overview_crossplot(
  toydata,
  id = ccode,
  time = year,
  cond1 = gdp,
  cond2 = population,
  threshold1 = 25000,
  threshold2 = 27000,
  color = TRUE,
  label = TRUE
)
```

Figure 4 shows where the units of observations (country-year) are located across both conditions (`gdp` and `population`).

In a next step, the user may want to compare two data sets. For this, we artificially reducing the internal data set to allow a comparison.

```
# Subset one data set for comparison
toydata2 <- toydata %>% dplyr::filter(year > 1992)

overview_overlap(
  dat1 = toydata,
  dat2 = toydata2,
  dat1_id = ccode,
  dat2_id = ccode,
  plot_type = "bar" # This is the default
)
```

Figure 5 shows that there are less observations for Rwanda (`RWA`) and the United Kingdom (`GBR`) in `toydata2` – which is true since we subset this data set and only kept observations for 1992 and newer.

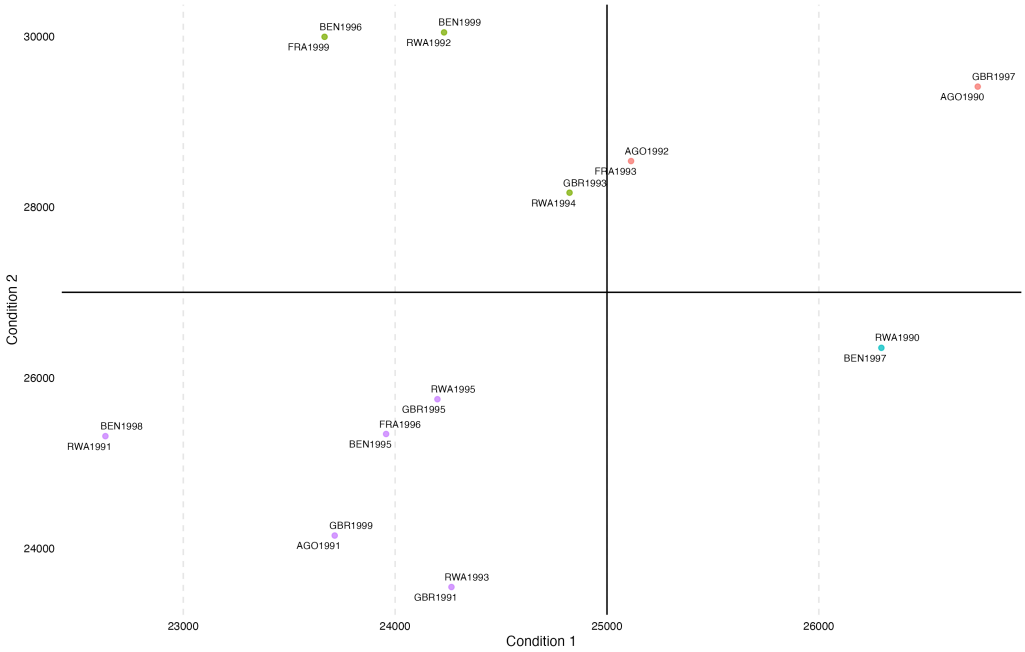

Figure 4: Visualization of 'overview_crossplot'

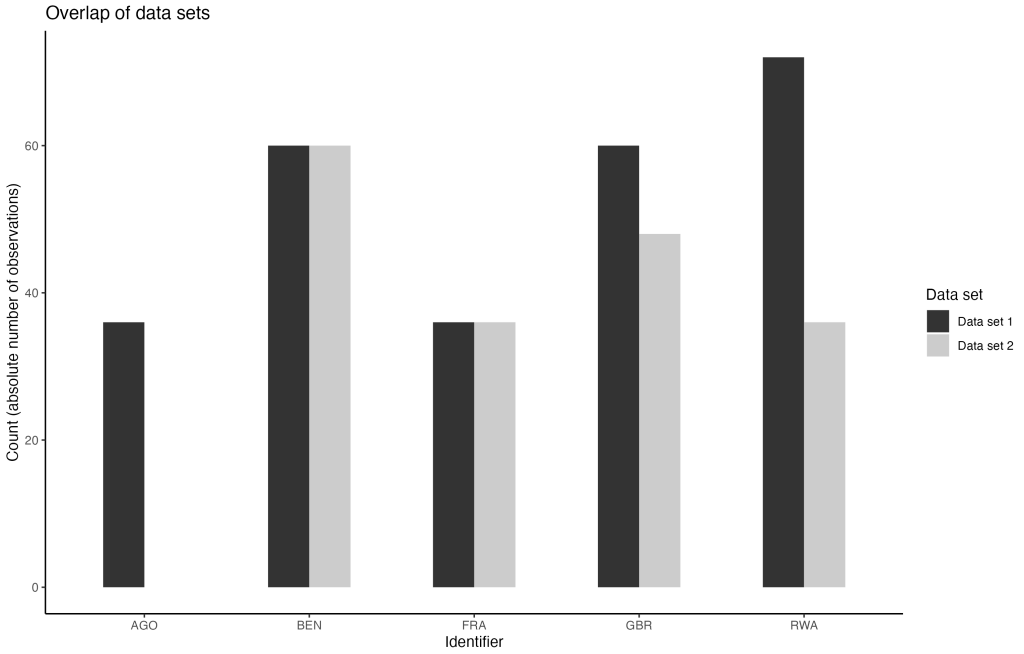

Figure 5: Visualization of 'overview_overlap'

Availability

overviewR is available on CRAN and the development version is available on Github. We also provide a cheatsheet to quickly get an idea of its key functionalities and describe possible workflows on the package website.

Acknowledgements

We would like to thank the excellent R community that enabled the creation of overviewR by providing free resources on package development.

References

Arel-Bundock, Vincent, CJ Yetman, Nils Enevoldsen, and Samuel Meichtry. 2022. *Convert Country Names and Country Codes* (version v1.4.0). https://CRAN.R-project.org/package=countrycode.

Comtois, Dominic. 2022. *Summarytools* (version v1.0.1). https://CRAN.R-project.org/package=summarytools.

Hammerschmidt, Dennis. 2021. "Moving Beyond Votes: Estimating and Analyzing State Relations Using Natural Language Processing, Complex Networks, and Machine Learning." PhD thesis, University of Mannheim.

Hammerschmidt, Dennis, Cosima Meyer, and Anne Pintsch. 2021. "Foreign Aid in Times of Populism: The Influence of Populist Radical Right Parties on the Official Development Assistance of OECD Countries." *Cambridge Review of International Affairs*, 1–22. https://doi.org/https://doi.org/10.1080/09557571.2021.1980498.

Linlin, Yan. 2021. *Draw Venn Diagram by 'Ggplot2'* (version v0.1.9). https://CRAN.R-project.org/package=ggvenn.

Meyer, Cosima. 2021. "Power Struggle and Spark of Hope: The Political Elite and Post-Civil War Politics." PhD thesis, University of Mannheim.

Meyer, Cosima, and Britt Bolin. 2022. "Power in the Post-Civil War Period: The Effect of Armed Conflict and Gender Quotas on Women in Political Leadership Positions." *Journal of Global Security Studies* 7 (4). https://doi.org/https://doi.org/10.1093/jogss/ogac009.

Meyer, Cosima, and Dennis Hammerschmidt. 2022. *overviewR* (version v0.0.10). https://CRAN.R-project.org/package=overviewR.

Putatunda, Sayan, Dayananda Ubrangala, Kiran Rama, and Ravi Kondapalli. 2019. "SmartEDA: An r Package for Automated Exploratory Data Analysis." *Journal of Open Source Software* 4 (41): 1509. https://doi.org/10.21105/joss.01509.

Ryu, Choonghyun. 2022. *Dlookr: Tools for Data Diagnosis, Exploration, Transformation* (version v0.6.0). https://CRAN.R-project.org/package=dlookr.

Staniak, Mateusz, and Przemysław Biecek. 2019. "The Landscape of R Packages for Automated Exploratory Data Analysis." *The R Journal* 11 (2): 347–69. https://doi.org/10.32614/RJ-2019-033.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and RStudio. 2022. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (version v3.3.6). https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and RStudio. 2022. *Dplyr: A Grammar of Data Manipulation* (version v1.0.9). https://CRAN.R-project.org/package=dplyr.