Smallset Timelines with smallsets: : **CHEAT SHEET**

smallsets is a tool for visually documenting and communicating data preprocessing decisions. It builds a Smallset Timeline figure [3] based on preprocessing code in an R, R Markdown, or Python script. Users must first add structured comments, with building instructions, to the preprocessing code.

Structured comments

Structure

`# smallsets` *`instruction data`* `caption[`*`text`*`]caption`

Mandatory instructions

`start`	start code tracking and take first snapshot
`end`	end code tracking and take final snapshot

Optional instructions

`snap`	take intermediate snapshot after next line
`resume`	add resume marker after next line

Main functions

`Smallset_Timeline(data, code, ...)`
builds a Smallset Timeline

`sets_sizing()`
for adjusting sizing parameters, including column names, caption text, snapshot data, and legend items

`sets_spacing()`
for adjusting spacing parameters, including caption space, column name rotation, and number of figure rows

`sets_labelling()`
for adjusting the colours of the column names and snapshot data

Demo dataset and code

The smallsets package comes with example data and preprocessing code, which are used to illustrate how the package works, such as in the next section.

Synthetic dataset

`s_data` 100 observations and eight variables (C1-C8)

Preprocessing scripts

`s_data_preprocess.R`	basic preprocessing scenario in R
`s_data_preprocess.Rmd`	basic preprocessing scenario in R Markdown
`s_data_preprocess.py`	basic preprocessing scenario in Python
`s_data_preprocess_4.R`	includes additional snapshot
`s_data_preprocess_resume.R`	includes resume marker

Steps to build a Smallset Timeline

The demo dataset `s_data` and preprocessing code `s_data_preprocess.R` are used to illustrate the process.

Step 1

Add structured comments to the preprocessing code in your R, R Markdown, or Python script, specifying snapshot points and captions.

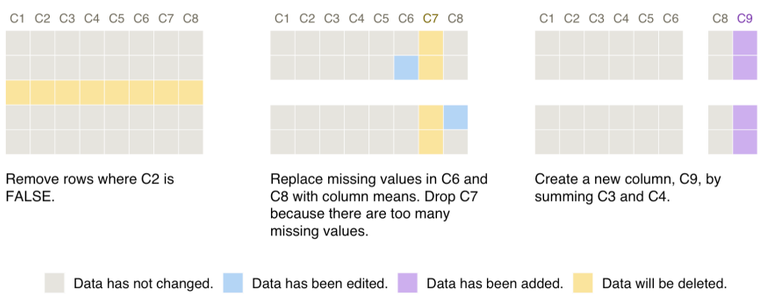
File: `s_data_preprocess.R`

```
1  # smallsets start s_data caption[Remove rows where C2 is FALSE.]caption
2  s_data <- s_data[s_data$C2 == TRUE,]
3
4  s_data$C6[is.na(s_data$C6)] <- mean(s_data$C6, na.rm = TRUE)
5  # smallsets snap s_data caption[Replace missing values in C6 and C8 with column
6  # means. Drop C7 because there are too many missing values.]caption
7  s_data$C8[is.na(s_data$C8)] <- mean(s_data$C8, na.rm = TRUE)
8  s_data$C7 <- NULL
9
10 s_data$C9 <- s_data$C3 + s_data$C4
11 # smallsets end s_data caption[Create a new column, C9, by summing C3 and
12 # C4.]caption
```

Step 2

Run the main smallsets command to build a Smallset Timeline for your dataset and preprocessing code.

```
Smallset_Timeline(data=s_data, code="s_data_preprocess.R")
```



Smallset selection

To select the small number of rows from the original dataset used in the visualisation, you can use one of three selection methods available in the `Smallset_Timeline()` command.

`rowCount` number of Smallset rows (5-15)

`rowSelect` Smallset row selection method

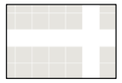
- `= 1` → *coverage* model (Gurobi required)
- `= 2` → *coverage* + *variety* model (Gurobi required)
- `= NULL` → random sampling

Warning. This method has long runtimes for large datasets. See [2] for workarounds.

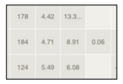
Customisation

To customise the information in a Smallset Timeline and its appearance, you can set different parameters in the `Smallset_Timeline()` command. See [1] for the complete list of parameters.

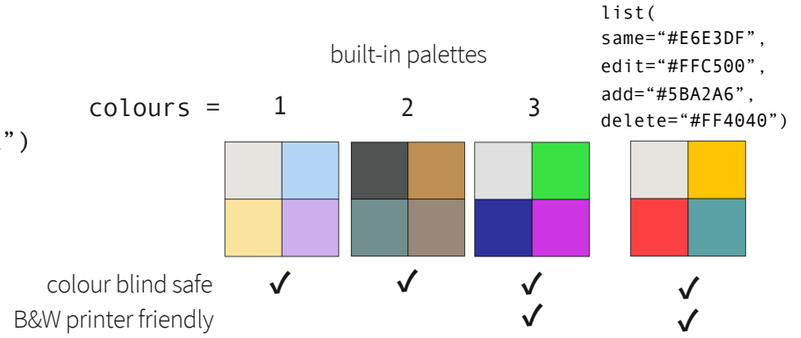
`printedData = TRUE`
show the data values in Smallset snapshots



`ghostData = TRUE`
plot blank rows/columns after data deletions



`missingDataTints = TRUE`
use colour tints to highlight missing data



References

[1] CRAN reference manual
cran.r-project.org/web/packages/smallsets/smallsets.pdf

[2] smallsets User Guide
lydialucchesi.github.io/smallsets/articles/smallsets.html
included in the package : `vignette("smallsets")`

[3] Smallset Timelines: A Visual Representation of Data Preprocessing Decisions
paper providing a detailed discussion of Smallset Timelines, the Smallset selection optimisation problems, and two case studies with example Smallset Timelines
doi.org/10.1145/3531146.3533175