Package 'sjmisc'

February 3, 2018

Type Package

Encoding UTF-8

Title Data and Variable Transformation Functions

Version 2.7.0

Date 2018-02-03

Author Daniel Lüdecke <d.luedecke@uke.de>

Maintainer Daniel Lüdecke <d.luedecke@uke.de>

Description Collection of miscellaneous utility functions, supporting data transformation tasks like recoding, dichotomizing or grouping variables, setting and replacing missing values. The data transformation functions also support labelled data, and all integrate seamlessly into a 'tidyverse'-workflow.

License GPL-3

Depends R (>= 3.2), stats, utils

Imports broom (>= 0.4.2), cli, crayon, dplyr (>= 0.7.1), haven (>= 1.0.0), magrittr, pillar, psych, purrr, rlang, sjlabelled (>= 1.0.7), stringdist (>= 0.9.4), stringr (>= 1.2.0), tibble (>= 1.4.1), tidyr (>= 0.7.0), tidyselect

Suggests ggplot2, graphics, Hmisc, mice, sjPlot (>= 2.4.0), sjstats (>= 0.13.0), knitr, rmarkdown

URL https://github.com/strengejacke/sjmisc

BugReports https://github.com/strengejacke/sjmisc/issues

RoxygenNote 6.0.1

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2018-02-03 22:35:29 UTC

R topics documented:

`sjmisc-package` *Data and Variable Transformation Functions*

Description

Purpose of this package

Collection of miscellaneous utility functions, supporting data transformation tasks like recoding, dichotomizing or grouping variables, setting and replacing missing values. The data transformation functions also support labelled data, and all integrate seamlessly into a 'tidyverse'-workflow.

Design philosophy - consistent api

The design of 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.

There are two types of function designs:

transformation/recoding functions Functions like `rec()` or `dicho()`, which transform or recode variables, typically do *not* return the complete data frame that was given as first argument, but only the transformed and recoded variables specified in the `...`-ellipses argument. The variables usually get a suffix, so you can bind these variables as new columns to a data frame.

coercing/converting functions Functions like `to_factor()` or `to_label()`, which convert variables into other types or add additional information like variable or value labels as attribute, typically return the complete data frame that was given as first argument. The variables specified in the `...`-ellipses argument are converted, all other variables remain unchanged.

Author(s)

Daniel Lüdecke <d.luedecke@uke.de>

`add_columns` *Add or replace data frame columns*

Description

`add_columns()` combines two or more data frames, but unlike `cbind` or `bind_cols`, this function binds `data` as last columns of a data frame.

`replace_columns()` replaces all columns in `data` with identically named columns in `...`, and adds remaining (non-duplicated) columns from `...` to `data`.

Usage

```
add_columns(data, ..., replace = TRUE)

replace_columns(data, ..., add.unique = TRUE)
```

Arguments

data	A data frame. For add_columns(), will be bound after data frames specified in .... For replace_columns(), duplicated columns in data will be replaced by columns in ....
...	More data frames to combine, resp. more data frames with columns that should replace columns in data.
replace	Logical, if TRUE (default), columns in ... with identical names in data will replace the columns in data. The order of columns after replacing is preserved.
add.unique	Logical, if TRUE (default), remaining columns in ... that did not replace any column in data, are appended as new columns to data.

Value

For add_columns(), a data frame, where columns of data are appended after columns of ....

For replace_columns(), a data frame where columns in data will be replaced by identically named columns in ..., and remaining columns from ... will be appended to data (if add.unique = TRUE).

Note

For add_columns(), by default, columns in data with identical names like columns in one of the data frames in ... will be dropped (i.e. variables with identical names in ... will replace existing variables in data). Use replace = FALSE to keep all columns. Identical column names will then be renamed, to ensure unique column names (which happens by default when using bind_cols). When replacing columns, replaced columns are not added to the end of the data frame. Rather, the original order of columns will be preserved.

Examples

```
data(efc)
d1 <- efc[, 1:3]
d2 <- efc[, 4:6]

library(dplyr)
head(bind_cols(d1, d2))
add_columns(d1, d2)

d1 <- efc[, 1:3]
d2 <- efc[, 2:6]

add_columns(d1, d2, replace = TRUE)
add_columns(d1, d2, replace = FALSE)
```

```
# use case: we take the original data frame, select specific
# variables and do some transformations or recodings
# (standardization in this example) and add the new, transformed
# variables *to the end* of the original data frame
efc %>%
  select(e17age, c160age) %>%
  std() %>%
  add_columns(efc)

# new variables with same name will overwrite old variables
# in "efc". order of columns is not changed.
efc %>%
  select(e16sex, e42dep) %>%
  to_factor() %>%
  add_columns(efc)

# keep both old and new variables, automatically
# rename variables with identical name
efc %>%
  select(e16sex, e42dep) %>%
  to_factor() %>%
  add_columns(efc, replace = FALSE)

# create sample data frames
d1 <- efc[, 1:10]
d2 <- efc[, 2:3]
d3 <- efc[, 7:8]
d4 <- efc[, 10:12]

# show original
head(d1)

# slightly change variables, to see effect
d2 <- to_label(d2)
d3 <- to_label(d3)

# replace duplicated columns, append remaining
replace_columns(d1, d2, d3, d4)

# replace duplicated columns, omit remaining
replace_columns(d1, d2, d3, d4, add.unique = FALSE)
```

`all_na` *Check if vector only has NA values*

Description

Check if all values in a vector are `NA`.

Usage

```
all_na(x)
```

Arguments

x	A vector or data frame.

Value

Logical, TRUE if x has only NA values, FALSE if x has at least one non-missing value.

Examples

```
x <- c(NA, NA, NA)
y <- c(1, NA, NA)

all_na(x)
all_na(y)
all_na(data.frame(x, y))
```

---

big_mark *Formats large numbers with big marks*

Description

Formats large numbers with big marks

Usage

```
big_mark(x, big.mark = ",", ...)
```

Arguments

x	A vector or data frame. All numeric inputs (including numeric character) vectors) will be prettified.
big.mark	Character, used as mark between every 3 decimals before the decimal point.
...	Other arguments passed down to the prettyNum-function.

Value

A prettified x as character, with big marks.

Examples

```
# simple big mark
big_mark(1234567)

# big marks for several values at once, mixed numeric and character
big_mark(c(1234567, "55443322"))

# pre-defined width of character output
big_mark(c(1234567, 55443322), width = 15)
```

`count_na` *Frequency table of tagged NA values*

Description

This method counts tagged NA values (see `tagged_na`) in a vector and prints a frequency table of counted tagged NAs.

Usage

```
count_na(x, ...)
```

Arguments

`x`	A vector or data frame.
`...`	Optional, unquoted names of variables that should be selected for further processing. Required, if `x` is a data frame (and no vector) and only selected variables from `x` should be processed. You may also use functions like `:` or tidyselect's `select_helpers`. See 'Examples' or package-vignette.

Value

A data frame with counted tagged NA values.

Examples

```
library(haven)
x <- labelled(
  x = c(1:3, tagged_na("a", "c", "z"),
        4:1, tagged_na("a", "a", "c"),
        1:3, tagged_na("z", "c", "c"),
        1:4, tagged_na("a", "c", "z")),
  labels = c("Agreement" = 1, "Disagreement" = 4,
             "First" = tagged_na("c"), "Refused" = tagged_na("a"),
             "Not home" = tagged_na("z"))
)
count_na(x)
```

```
y <- labelled(
  x = c(1:3, tagged_na("e", "d", "f"),
        4:1, tagged_na("f", "f", "d"),
        1:3, tagged_na("f", "d", "d"),
        1:4, tagged_na("f", "d", "f")),
  label = c("Agreement" = 1, "Disagreement" = 4, "An E" = tagged_na("e"),
            "A D" = tagged_na("d"), "The eff" = tagged_na("f"))
)

# create data frame
library(tibble)
dat <- tibble(x, y)

# possible count()-function calls
count_na(dat)
count_na(dat$x)
count_na(dat, x)
count_na(dat, x, y)
```

---

descr *Basic descriptive statistics*

Description

This function wraps the describe-function and prints a basic descriptive statistic, including variable labels.

Usage

```
descr(x, ..., max.length = NULL, out = c("txt", "viewer", "browser"))
```

Arguments

x	A vector or a data frame. May also be a grouped data frame (see 'Note' and 'Examples').
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
max.length	Numeric, indicating the maximum length of variable labels in the output. If variable names are longer than max.length, they will be shortened to the last whole word within the first max.length chars.
out	Character vector, indicating whether the results should be printed to console (out = "txt") or as HTML-table in the viewer-pane (out = "viewer") or browser (out = "browser").

Value

A data frame with basic descriptive statistics, derived from the describe-function. The additional column NA.prc informs about the percentage of missing values in a variable.

Note

data may also be a grouped data frame (see group_by) with up to two grouping variables. Descriptive tables are created for each subgroup then.

Examples

```
data(efc)
descr(efc, e17age, c160age)

library(dplyr)
efc %>% select(e42dep, e15relat, c172code) %>% descr()

# with grouped data frames
efc %>%
  group_by(e16sex) %>%
  select(e16sex, e42dep, e15relat, c172code) %>%
  descr()

# you can select variables also inside 'descr()'
efc %>%
  group_by(e16sex, c172code) %>%
  descr(e16sex, c172code, e17age, c160age)

# or even use select-helpers
descr(efc, contains("cop"), max.length = 20)
```

---

dicho *Dichotomize variables*

Description

Dichotomizes variables into dummy variables (0/1). Dichotomization is either done by median, mean or a specific value (see dich.by).

Usage

```
dicho(x, ..., dich.by = "median", as.num = FALSE, var.label = NULL,
  val.labels = NULL, append = TRUE, suffix = "_d")
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
dich.by	Indicates the split criterion where a variable is dichotomized. Must be one of the following values (may be abbreviated): "median" **or** "md" by default, x is split into two groups at the median. "mean" **or** "m" splits x into two groups at the mean of x. **numeric value** splits x into two groups at the specific value. Note that the value is inclusive, i.e. dich.by = 10 will split x into one group with values from lowest to 10 and another group with values greater than 10.
as.num	Logical, if TRUE, return value will be numeric, not a factor.
var.label	Optional string, to set variable label attribute for the returned variable (see vignette Labelled Data and the sjlabelled-Package). If NULL (default), variable label attribute of x will be used (if present). If empty, variable label attributes will be removed.
val.labels	Optional character vector (of length two), to set value label attributes of dichotomized variable (see set_labels). If NULL (default), no value labels will be set.
append	Logical, if TRUE (the default) and x is a data frame, x including the new variables as additional columns is returned; if FALSE, only the new variables are returned.
suffix	String value, will be appended to variable (column) names of x, if x is a data frame. If x is not a data frame, this argument will be ignored. The default value to suffix column names in a data frame depends on the function call: • recoded variables (rec()) will be suffixed with "_r" • recoded variables (recode_to()) will be suffixed with "_r0" • dichotomized variables (dicho()) will be suffixed with "_d" • grouped variables (split_var()) will be suffixed with "_g" • grouped variables (group_var()) will be suffixed with "_gr" • standardized variables (std()) will be suffixed with "_z" • centered variables (center()) will be suffixed with "_c"

Details

dicho() also works on grouped data frames (see group_by). In this case, dichotomization is applied to the subsets of variables in x. See 'Examples'.

Value

x, dichotomized. If x is a data frame, only the dichotomized variables will be returned.

Note

Variable label attributes are preserved (unless changed via `var.label`-argument).

Examples

```
data(efc)
summary(efc$c12hour)
# split at median
table(dicho(efc$c12hour))
# split at mean
table(dicho(efc$c12hour, dich.by = "mean"))
# split between value lowest to 30, and above 30
table(dicho(efc$c12hour, dich.by = 30))

# sample data frame, values from 1-4
head(efc[, 6:10])

# dichtomized values (1 to 2 = 0, 3 to 4 = 1)
library(dplyr)
efc %>%
  select(6:10) %>%
  dicho(dich.by = 2) %>%
  head()

# dichtomize several variables in a data frame
dicho(efc, c12hour, e17age, c160age, append = FALSE)

# dichotomize and set labels
frq(dicho(
  efc, e42dep,
  var.label = "Dependency (dichotomized)",
  val.labels = c("lower", "higher"),
  append = FALSE
))

# works also with gouped data frames
mtcars %>%
  dicho(disp, append = FALSE) %>%
  table()

mtcars %>%
  group_by(cyl) %>%
  dicho(disp, append = FALSE) %>%
  table()

# dichotomizing grouped data frames leads to different
# results for a dichotomized variable, because the split
# value is different for each group.
# compare:
mtcars %>%
  group_by(cyl) %>%
  summarise(median = median(disp))
```

```
median(mtcars$disp)
```

---

efc *Sample dataset from the EUROFAMCARE project*

Description

A SPSS sample data set, imported with the read_spss function.

Examples

```
# Attach EFC-data
data(efc)

# Show structure
str(efc)

# show first rows
head(efc)
```

---

empty_cols *Return or remove variables or observations that are completely missing*

Description

These functions check which rows or columns of a data frame completely contain missing values, i.e. which observations or variables completely have missing values, and either 1) returns their indices; or 2) removes them from the data frame.

Usage

```
empty_cols(x)

empty_rows(x)

remove_empty_cols(x)

remove_empty_rows(x)
```

Arguments

x	A data frame.

Value

For `empty_cols` and `empty_rows`, a numeric (named) vector with row or column indices of those variables that completely have missing values.

For `remove_empty_cols` and `remove_empty_rows`, a data frame with "empty" columns or rows removed.

Examples

```
tmp <- data.frame(a = c(1, 2, 3, NA, 5),
                  b = c(1, NA, 3, NA , 5),
                  c = c(NA, NA, NA, NA, NA),
                  d = c(1, NA, 3, NA, 5))

tmp

empty_cols(tmp)
empty_rows(tmp)

remove_empty_cols(tmp)
remove_empty_rows(tmp)
```

---

find_var *Find variable by name or label*

Description

This functions finds variables in a data frame, which variable names or variable (and value) label attribute match a specific pattern. Regular expression for the pattern is supported.

Usage

```
find_var(data, pattern, ignore.case = TRUE, search = c("name_label",
  "name_value", "label_value", "name", "label", "value", "all"),
  out = c("table", "df", "index"), fuzzy = FALSE, as.df, as.varlab)
```

Arguments

`data`	A data frame.
`pattern`	Character string to be matched in `data`. May also be a character vector of length > 1 (see 'Examples'). `pattern` is searched for in column names and variable label attributes of `data` (see `get_label`). `pattern` might also be a regular-expression object, as returned by `regex`, or any of **stringr**'s supported `modifiers`.
`ignore.case`	Logical, whether matching should be case sensitive or not.

search — Character string, indicating where pattern is sought. Use one of following options:

- "name_label" The default, searches for pattern in variable names and variable labels.
- "name_value" Searches for pattern in variable names and value labels.
- "label_value" Searches for pattern in variable and value labels.
- "name" Searches for pattern in variable names.
- "label" Searches for pattern in variable labels
- "value" Searches for pattern in value labels.
- "all" Searches for pattern in variable names, variable and value labels.

out — Output (return) format of the search results. May be abbreviated and must be one of:

- "table" A tabular overview (as data frame) with column indices, variable names and labels of matching variables.
- "df" A data frame with all matching variables.
- "index" A named vector with column indices of all matching variables.

fuzzy — Logical, if TRUE, "fuzzy matching" (partial and close distance matching) will be used to find pattern in data if no exact match was found. str_pos is used for fuzzy matching.

as.df — Deprecated, use out = "df" instead.

as.varlab — Deprecated, use out = "table" instead.

Details

This function searches for pattern in data's column names and - for labelled data - in all variable and value labels of data's variables (see get_label for details on variable labels and labelled data). Search is performed using the str_detect functions; hence, regular expressions are supported as well, by simply using pattern = stringr::regex(...).

Value

By default (i.e. out = "table", returns a tibble with three columns: column number, variable name and variable label. If out = "index", returns a named vector with column indices of matching variables (variable names are used as names-attribute); if out = "df", returns the matching variables as tibble.

Examples

```
data(efc)

# find variables with "cop" in variable name
find_var(efc, "cop")

# return tibble with matching variables
find_var(efc, "cop", out = "df")

# or return column numbers
```

```
find_var(efc, "cop", out = "index")

# find variables with "dependency" in names and variable labels
library(sjlabelled)
find_var(efc, "dependency")
get_label(efc$e42dep)

# find variables with "level" in names and value labels
res <- find_var(efc, "level", search = "name_value", out = "df")
res
get_labels(res, attr.only = FALSE)

# use sjPlot::view_df() to view results
## Not run:
library(sjPlot)
view_df(res)
## End(Not run)
```

`flat_table` *Flat (proportional) tables*

Description

This function creates a labelled flat table or flat proportional (marginal) table.

Usage

```
flat_table(data, ..., margin = c("counts", "cell", "row", "col"),
  digits = 2, show.values = FALSE)
```

Arguments

`data`	A data frame. May also be a grouped data frame (see 'Note' and 'Examples').
`...`	One or more variables of `data` that should be printed as table.
`margin`	Specify the table margin that should be computed for proportional tables. By default, counts are printed. Use `margin = "cell"`, `margin = "col"` or `margin = "row"` to print cell, column or row percentages of the table margins.
`digits`	Numeric; for proportional tables, `digits` indicates the number of decimal places.
`show.values`	Logical, if `TRUE`, value labels are prefixed by the associated value.

Value

An object of class `ftable`.

Note

`data` may also be a grouped data frame (see `group_by`) with up to two grouping variables. Cross tables are created for each subgroup then.

See Also

frq for simple frequency table of labelled vectors.

Examples

```
data(efc)

# flat table with counts
flat_table(efc, e42dep, c172code, e16sex)

# flat table with proportions
flat_table(efc, e42dep, c172code, e16sex, margin = "row")

# flat table from grouped data frame. You need to select
# the grouping variables and at least two more variables for
# cross tabulation.
library(dplyr)
efc %>%
  group_by(e16sex) %>%
  select(e16sex, c172code, e42dep) %>%
  flat_table()

efc %>%
  group_by(e16sex, e42dep) %>%
  select(e16sex, e42dep, c172code, n4pstu) %>%
  flat_table()

# now it gets weird...
efc %>%
  group_by(e16sex, e42dep) %>%
  select(e16sex, e42dep, c172code, n4pstu, c161sex) %>%
  flat_table()
```

---

frq *Frequencies of labelled variables*

Description

This function returns a frequency table of labelled vectors, as data frame.

Usage

```
frq(x, ..., sort.frq = c("none", "asc", "desc"), weight.by = NULL,
  auto.grp = NULL, show.strings = TRUE, grp.strings = NULL,
  out = c("txt", "viewer", "browser"))
```

Arguments

x	A vector or a data frame. May also be a grouped data frame (see 'Note' and 'Examples').
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
sort.frq	Determines whether categories should be sorted according to their frequencies or not. Default is "none", so categories are not sorted by frequency. Use "asc" or "desc" for sorting categories ascending or descending order.
weight.by	Vector of weights that will be applied to weight all observations. Must be a vector of same length as the input vector. Default is NULL, so no weights are used.
auto.grp	Numeric value, indicating the minimum amount of unique values in a variable, at which automatic grouping into smaller units is done (see group_var). Default value for auto.group is NULL, i.e. auto-grouping is off.
show.strings	Logical, if TRUE, frequency tables for character vectors will not be printed. This is useful when printing frequency tables of all variables from a data frame, and due to computational reasons character vectors should not be printed.
grp.strings	Numeric, if not NULL, groups string values in character vectors, based on their similarity. The similarity is estimated with the **stringdist**-package. See group_str for details on grouping, and that function's maxdist-argument to get more details on the distance of strings to be treated as equal.
out	Character vector, indicating whether the results should be printed to console (out = "txt") or as HTML-table in the viewer-pane (out = "viewer") or browser (out = "browser").

Value

A list of data frames with values, value labels, frequencies, raw, valid and cumulative percentages of x.

Note

x may also be a grouped data frame (see group_by) with up to two grouping variables. Frequency tables are created for each subgroup then.

See Also

flat_table for labelled (proportional) tables.

Examples

```
library(haven)
# create labelled integer
x <- labelled(
  c(1, 2, 1, 3, 4, 1),
```

```
  c(Male = 1, Female = 2, Refused = 3, "N/A" = 4)
)
frq(x)

x <- labelled(
  c(1:3, tagged_na("a", "c", "z"), 4:1, 2:3),
  c("Agreement" = 1, "Disagreement" = 4, "First" = tagged_na("c"),
    "Refused" = tagged_na("a"), "Not home" = tagged_na("z"))
)
frq(x)

# in a pipe
data(efc)
library(dplyr)
efc %>%
  select(e42dep, e15relat, c172code) %>%
  frq()

# or:
# frq(efc, e42dep, e15relat, c172code)

# with grouped data frames, in a pipe
efc %>%
  group_by(e16sex, c172code) %>%
  frq(e16sex, c172code, e42dep)

# with select-helpers: all variables from the COPE-Index
# (which all have a "cop" in their name)
frq(efc, contains("cop"))

# all variables from column "c161sex" to column "c175empl"
frq(efc, c161sex:c175empl)

# for non-labelled data, variable name is printed,
# and "label" column is removed from output
data(iris)
frq(iris, Species)

# group variables with large range
frq(efc, c160age)
frq(efc, c160age, auto.grp = 5)

# group string values
## Not run:
dummy <- efc %>% dplyr::select(3)
dummy$words <- sample(
  c("Hello", "Helo", "Hole", "Apple", "Ape",
    "New", "Old", "System", "Systemic"),
  size = nrow(dummy),
  replace = TRUE
)

frq(dummy)
```

```
frq(dummy, grp.strings = 2)
## End(Not run)
```

---

group_str *Group near elements of string vectors*

Description

This function groups elements of a string vector (character or string variable) according to the element's distance ('similatiry'). The more similar two string elements are, the higher is the chance to be combined into a group.

Usage

```
group_str(strings, maxdist = 2, method = "lv", strict = FALSE,
  trim.whitespace = TRUE, remove.empty = TRUE, showProgressBar = FALSE)
```

Arguments

strings	Character vector with string elements.
maxdist	Maximum distance between two string elements, which is allowed to treat two elements as similar or equal.
method	Method for distance calculation. The default is "lv". See stringdist for details.
strict	Logical; if TRUE, value matching is more strictly. See 'Examples'.
trim.whitespace	Logical; if TRUE (default), leading and trailing white spaces will be removed from string values.
remove.empty	Logical; if TRUE (default), empty string values will be removed from the character vector strings.
showProgressBar	Logical; if TRUE, the progress bar is displayed when computing the distance matrix. Default in FALSE, hence the bar is hidden.

Value

A character vector where similar string elements (values) are recoded into a new, single value. The return value is of same length as strings, i.e. grouped elements appear multiple times, so the count for each grouped string is still avaiable (see 'Examples').

See Also

str_pos

Examples

```
oldstring <- c("Hello", "Helo", "Hole", "Apple",
               "Ape", "New", "Old", "System", "Systemic")
newstring <- group_str(oldstring)

# see result
newstring

# count for each groups
table(newstring)

# print table to compare original and grouped string
frq(oldstring)
frq(newstring)

# larger groups
newstring <- group_str(oldstring, maxdist = 3)
frq(oldstring)
frq(newstring)

# be more strict with matching pairs
newstring <- group_str(oldstring, maxdist = 3, strict = TRUE)
frq(oldstring)
frq(newstring)
```

---

group_var *Recode numeric variables into equal-ranged groups*

Description

Recode numeric variables into equal ranged, grouped factors, i.e. a variable is cut into a smaller number of groups, where each group has the same value range, and create the related value labels.

Usage

```
group_var(x, ..., size = 5, as.num = TRUE, right.interval = FALSE,
  n = 30, append = TRUE, suffix = "_gr")

group_labels(x, ..., size = 5, right.interval = FALSE, n = 30)
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.

size	Numeric; group-size, i.e. the range for grouping. By default, for each 5 categories of x a new group is defined, i.e. size = 5. Use size = "auto" to automatically resize a variable into a maximum of 30 groups (which is the ggplot-default grouping when plotting histograms). Use n to determine the amount of groups.
as.num	Logical, if TRUE, return value will be numeric, not a factor.
right.interval	Logical; if TRUE, grouping starts with the lower bound of size. See 'Details'.
n	Sets the maximum number of groups that are defined when auto-grouping is on (size = "auto"). Default is 30. If size is not set to "auto", this argument will be ignored.
append	Logical, if TRUE (the default) and x is a data frame, x including the new variables as additional columns is returned; if FALSE, only the new variables are returned.
suffix	String value, will be appended to variable (column) names of x, if x is a data frame. If x is not a data frame, this argument will be ignored. The default value to suffix column names in a data frame depends on the function call: • recoded variables (rec()) will be suffixed with "_r" • recoded variables (recode_to()) will be suffixed with "_r0" • dichotomized variables (dicho()) will be suffixed with "_d" • grouped variables (split_var()) will be suffixed with "_g" • grouped variables (group_var()) will be suffixed with "_gr" • standardized variables (std()) will be suffixed with "_z" • centered variables (center()) will be suffixed with "_c"

Details

If size is set to a specific value, the variable is recoded into several groups, where each group has a maximum range of size. Hence, the amount of groups differ depending on the range of x.

If size = "auto", the variable is recoded into a maximum of n groups. Hence, independent from the range of x, always the same amount of groups are created, so the range within each group differs (depending on x's range).

right.interval determins which boundary values to include when grouping is done. If TRUE, grouping starts with the **lower bound** of size. For example, having a variable ranging from 50 to 80, groups cover the ranges from 50-54, 55-59, 60-64 etc. If FALSE (default), grouping starts with the upper bound of size. In this case, groups cover the ranges from 46-50, 51-55, 56-60, 61-65 etc. **Note:** This will cover a range from 46-50 as first group, even if values from 46 to 49 are not present. See 'Examples'.

If you want to split a variable into a certain amount of equal sized groups (instead of having groups where values have all the same range), use the split_var function!

group_var() also works on grouped data frames (see group_by). In this case, grouping is applied to the subsets of variables in x. See 'Examples'.

Value

- For group_var, a grouped variable, either as numeric or as factor (see paramter as.num). If x is a data frame, only the grouped variables will be returned.
- For group_label, a string vector or a list of string vectors containing labels based on the grouped categories of x, formatted as "from lower bound to upper bound", e.g. "10-19" "20-29" "30-39" etc. See 'Examples'.

Note

Variable label attributes (see, for instance, set_label) are preserved. Usually you should use the same values for size and right.interval in group_label() as used in the group_var function if you want matching labels for the related recoded variable.

See Also

split_var to split variables into equal sized groups, group_str for grouping string vectors or rec_pattern and rec for another convenbient way of recoding variables into smaller groups.

Examples

```
age <- abs(round(rnorm(100, 65, 20)))
age.grp <- group_var(age, size = 10)
hist(age)
hist(age.grp)

age.grpvar <- group_labels(age, size = 10)
table(age.grp)
print(age.grpvar)

# histogram with EUROFAMCARE sample dataset
# variable not grouped
library(sjlabelled)
data(efc)
hist(efc$e17age, main = get_label(efc$e17age))

# bar plot with EUROFAMCARE sample dataset
# grouped variable
ageGrp <- group_var(efc$e17age)
ageGrpLab <- group_labels(efc$e17age)
barplot(table(ageGrp), main = get_label(efc$e17age), names.arg = ageGrpLab)

# within a pipe-chain
library(dplyr)
efc %>%
  select(e17age, c12hour, c160age) %>%
  group_var(size = 20)

# create vector with values from 50 to 80
dummy <- round(runif(200, 50, 80))
# labels with grouping starting at lower bound
group_labels(dummy)
```

```
# labels with grouping startint at upper bound
group_labels(dummy, right.interval = TRUE)

# works also with gouped data frames
mtcars %>%
  group_var(disp, size = 4, append = FALSE) %>%
  table()

mtcars %>%
  group_by(cyl) %>%
  group_var(disp, size = 4, append = FALSE) %>%
  table()
```

`is_crossed` *Check whether two factors are crossed or nested*

Description

These functions checks whether two factors are crossed or nested, i.e. if each level of one factor occurs in combination with each level of the other factor (`is_crossed()`) resp. if each category of the first factor co-occurs with only one category of the other (`is_nested()`).

Usage

```
is_crossed(f1, f2)

is_nested(f1, f2)
```

Arguments

`f1`	Numeric vector or `factor`.
`f2`	Numeric vector or `factor`.

Value

Logical. For `is_crossed()`, `TRUE` if factors are crossed, `FALSE` otherwise. For `nested()`, `TRUE` if factors are nested, `FALSE` otherwise.

Note

If factors are nested, a message is displayed to tell whether `f1` is nested within `f2` or vice versa.

References

Grace, K. The Difference Between Crossed and Nested Factors. (web)

Examples

```
# crossed factors, each category of
# x appears in each category of y
x <- c(1,4,3,2,3,2,1,4)
y <- c(1,1,1,2,2,1,2,2)
# show distribution
table(x, y)
# check if crossed
is_crossed(x, y)

# not crossed factors
x <- c(1,4,3,2,3,2,1,4)
y <- c(1,1,1,2,1,1,2,2)
# show distribution
table(x, y)
# check if crossed
is_crossed(x, y)


# nested factors, each category of
# x appears in one category of y
x <- c(1,2,3,4,5,6,7,8,9)
y <- c(1,1,1,2,2,2,3,3,3)
# show distribution
table(x, y)
# check if nested
is_nested(x, y)
is_nested(y, x)

# not nested factors
x <- c(1,2,3,4,5,6,7,8,9,1,2)
y <- c(1,1,1,2,2,2,3,3,3,2,3)
# show distribution
table(x, y)
# check if nested
is_nested(x, y)
is_nested(y, x)
```

---

is_empty *Check whether string, list or vector is empty*

Description

This function checks whether a string or character vector (of length 1), a list or any vector (numeric, atomic) is empty or not.

Usage

```
is_empty(x, first.only = TRUE)
```

Arguments

x	String, character vector of length 1, list or vector.
first.only	Logical, if FALSE and x is a character vector, each element of x will be checked if empty. If TRUE, only the first element of x will be checked.

Value

Logical, TRUE if x is a character vector or string and is empty, TRUE if x is a vector or list and of length 0, FALSE otherwise.

Note

NULL- or NA-values are also considered as "empty" (see 'Examples') and will return TRUE.

Examples

```
x <- "test"
is_empty(x)

x <- ""
is_empty(x)

x <- NA
is_empty(x)

x <- NULL
is_empty(x)

# string is not empty
is_empty(" ")

# however, this trimmed string is
is_empty(trim(" "))

# numeric vector
x <- 1
is_empty(x)
x <- x[-1]
is_empty(x)

# check multiple elements of character vectors
is_empty(c("", "a"))
is_empty(c("", "a"), first.only = FALSE)
```

is_even *Check whether value is even or odd*

Description

Checks whether x is an even or odd number. Only accepts numeric vectors.

Usage

```
is_even(x)

is_odd(x)
```

Arguments

x	Numeric vector or single numeric value, or a data frame or list with such vectors.

Value

is_even() returns TRUE for each even value of x, FALSE for odd values. is_odd() returns TRUE for each odd value of x and FALSE for even values.

Examples

```
is_even(4)
is_even(5)
is_even(1:4)

is_odd(4)
is_odd(5)
is_odd(1:4)
```

is_float *Check if a variable is of (non-integer) double type or a whole number*

Description

is_float() checks whether an input vector or value is a numeric non-integer (double), depending on fractional parts of the value(s). is_whole() does the opposite and checks whether an input vector is a whole number (without fractional parts).

Usage

```
is_float(x)

is_whole(x)
```

Arguments

x	A value, vector or data frame.

Value

For `is_float()`, `TRUE` if `x` is a floating value (non-integer double), `FALSE` otherwise (also returns `FALSE` for character vectors and factors). For `is_whole()`, `TRUE` if `x` is a vector with whole numbers only, `FALSE` otherwise (returns `TRUE` for character vectors and factors).

Examples

```
data(mtcars)
data(iris)

is.double(4)
is_float(4)
is_float(4.2)
is_float(iris)

is_whole(4)
is_whole(4.2)
is_whole(mtcars)
```

---

is_num_fac *Check whether a factor has numeric levels only*

Description

This function checks whether a factor has only numeric or any non-numeric factor levels.

Usage

```
is_num_fac(x)
```

Arguments

x	A `factor`.

Value

Logical, `TRUE` if factor has numeric factor levels only, `FALSE` otherwise.

Examples

```
# numeric factor levels
f1 <- factor(c(NA, 1, 3, NA, 2, 4))
is_num_fac(f1)

# not completeley numeric factor levels
f2 <- factor(c(NA, "C", 1, 3, "A", NA, 2, 4))
is_num_fac(f2)

# not completeley numeric factor levels
f3 <- factor(c("Justus", "Bob", "Peter"))
is_num_fac(f3)
```

---

merge_df *Merge labelled data frames*

Description

Merges (full join) two (or more) data frames and preserve value and variable labels.

Usage

```
merge_df(x1, x2, ..., id = NULL)
```

Arguments

x1	First data frame to be merged.
x2	Second data frame to be merged.
...	More data frames to be merged.
id	Optional name for ID column that will be created to indicate the source data frames for appended rows.

Details

This function merges two data frames, where equal named columns will be joined together. Matching rows are not omitted, hence all rows from the data frames are preserved. This means that merge_df row-binds all data frames, even if these have different numbers of columns. Non-matching columns will be column-bound and filled with NA-values for rows in those data frames that do not have this column.

Value and variable labels are preserved. If matching columns have different value label attributes, attributes from first data frame will be used.

Value

A full joined data frame.

Examples

```
library(dplyr)
data(efc)
x1 <- efc %>% select(1:5) %>% slice(1:10)
x2 <- efc %>% select(3:7) %>% slice(11:20)

mydf <- merge_df(x1, x2)
mydf
str(mydf)

## Not run:
library(sjPlot)
view_df(mydf)
## End(Not run)

x3 <- efc %>% select(5:9) %>% slice(21:30)
x4 <- efc %>% select(11:14) %>% slice(31:40)

mydf <- merge_df(x1, x2, x3, x4, id = "subsets")
mydf
str(mydf)
```

`merge_imputations` *Merges multiple imputed data frames into a single data frame*

Description

This function merges multiple imputed data frames from `mids`-objects into a single data frame by computing the mean or selecting the most likely imputed value.

Usage

```
merge_imputations(dat, imp, ori = NULL, summary = c("none", "dens", "hist",
  "sd"), filter = NULL)
```

Arguments

`dat`	The data frame that was imputed and used as argument in the `mice`-function call.
`imp`	The `mids`-object with the imputed data frames from `dat`.
`ori`	Optional, if `ori` is specified, the imputed variables are appended to this data frame; else, a new data frame with the imputed variables is returned.
`summary`	After merging multiple imputed data, `summary` displays a graphical summary of the "quality" of the merged values, compared to the original imputed values.
	`"dens"` Creates a density plot, which shows the distribution of the mean of the imputed values for each variable at each observation. The larger the areas overlap, the better is the fit of the merged value compared to the imputed value.

`"hist"` Similar to `summary = "dens"`, however, mean and merged values are shown as histogram. Bins should have almost equal height for both groups (mean and merged).

`"sd"` Creates a dot plot, where data points indicate the standard deviation for all imputed values (y-axis) at each merged value (x-axis) for all imputed variables. The higher the standard deviation, the less precise is the imputation, and hence the merged value.

`filter` A character vector with variable names that should be plotted. All non-defined variables will not be shown in the plot.

Details

This method merges multiple imputations of variables into a single variable by computing the (rounded) mean of all imputed values of missing values. By this, each missing value is replaced by those values that have been imputed the most times.

`imp` must be a `mids`-object, which is returned by the `mice`-function of the **mice**-package. `merge_imputations` than creates a data frame for each imputed variable, by combining all imputations (as returned by the `complete`-function) of each variable, and computing the row means of this data frame. The mean value is then rounded for integer values (and not for numerical values with fractional part), which corresponds to the most frequent imputed value for a missing value. The original variable with missings is then copied and missing values are replaced by the most frequent imputed value.

Value

A data frame with (merged) imputed variables; or `ori` with appended imputed variables, if `ori` was specified. If `summary` is included, returns a list with the data frame `data` with (merged) imputed variables and some other summary information, which are required for the plot-output.

Note

Typically, further analyses are conducted on pooled results of multiple imputed data sets (see `pool`), however, sometimes (in social sciences) it is also feasible to compute the mean of multiple imputed variables (see *Burns et al. 2011*).

References

Burns RA, Butterworth P, Kiely KM, Bielak AAM, Luszcz MA, Mitchell P, et al. 2011. Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data. Journal of Clinical Epidemiology;64:787–93 doi: 10.1016/j.jclinepi.2010.10.011

Examples

```
library(mice)
imp <- mice(nhanes)

# return data frame with imputed variables
merge_imputations(nhanes, imp)

# append imputed variables to original data frame
```

```
merge_imputations(nhanes, imp, nhanes)

# show summary of quality of merging imputations
merge_imputations(nhanes, imp, summary = "dens", filter = c("chl", "hyp"))
```

---

rec *Recode variables*

Description

Recodes values of variables

Usage

```
rec(x, ..., rec, as.num = TRUE, var.label = NULL, val.labels = NULL,
  append = TRUE, suffix = "_r")
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
rec	String with recode pairs of old and new values. See 'Details' for examples. rec_pattern is a convenient function to create recode strings for grouping variables.
as.num	Logical, if TRUE, return value will be numeric, not a factor.
var.label	Optional string, to set variable label attribute for the returned variable (see vignette Labelled Data and the sjlabelled-Package). If NULL (default), variable label attribute of x will be used (if present). If empty, variable label attributes will be removed.
val.labels	Optional character vector, to set value label attributes of recoded variable (see vignette Labelled Data and the sjlabelled-Package). If NULL (default), no value labels will be set. Value labels can also be directly defined in the rec-syntax, see 'Details'.
append	Logical, if TRUE (the default) and x is a data frame, x including the new variables as additional columns is returned; if FALSE, only the new variables are returned.
suffix	String value, will be appended to variable (column) names of x, if x is a data frame. If x is not a data frame, this argument will be ignored. The default value to suffix column names in a data frame depends on the function call: • recoded variables (rec()) will be suffixed with "_r" • recoded variables (recode_to()) will be suffixed with "_r0" • dichotomized variables (dicho()) will be suffixed with "_d"

- grouped variables (split_var()) will be suffixed with "_g"
- grouped variables (group_var()) will be suffixed with "_gr"
- standardized variables (std()) will be suffixed with "_z"
- centered variables (center()) will be suffixed with "_c"

Details

The rec string has following syntax:

recode pairs each recode pair has to be separated by a ;, e.g. rec = "1=1; 2=4; 3=2; 4=3"

multiple values multiple old values that should be recoded into a new single value may be separated with comma, e.g. "1,2=1; 3,4=2"

value range a value range is indicated by a colon, e.g. "1:4=1; 5:8=2" (recodes all values from 1 to 4 into 1, and from 5 to 8 into 2)

value range for doubles for double vectors (with floating points), all values within the specified range are recoded; e.g. 1:2.5=1;2.6:3=2 recodes 1 to 2.5 into 1 and 2.6 to 3 into 2, but 2.55 would not be recoded (since it's not included in any of the specified ranges)

"min" **and** "max" minimum and maximum values are indicates by *min* (or *lo*) and *max* (or *hi*), e.g. "min:4=1; 5:max=2" (recodes all values from minimum values of x to 4 into 1, and from 5 to maximum values of x into 2)

"else" all other values, which have not been specified yet, are indicated by *else*, e.g. "3=1; 1=2; else=3" (recodes 3 into 1, 1 into 2 and all other values into 3)

"copy" the "else"-token can be combined with *copy*, indicating that all remaining, not yet recoded values should stay the same (are copied from the original value), e.g. "3=1; 1=2; else=copy" (recodes 3 into 1, 1 into 2 and all other values like 2, 4 or 5 etc. will not be recoded, but copied, see 'Examples')

NA**'s** NA values are allowed both as old and new value, e.g. "NA=1; 3:5=NA" (recodes all NA into 1, and all values from 3 to 5 into NA in the new variable)

"rev" "rev" is a special token that reverses the value order (see 'Examples')

direct value labelling value labels for new values can be assigned inside the recode pattern by writing the value label in square brackets after defining the new value in a recode pair, e.g. "15:30=1 [young aged]; 31:55=2 [middle aged]; 56:max=3 [old aged]". See 'Examples'.

Value

x with recoded categories. If x is a data frame, for append = TRUE, x including the recoded variables as new columns is returned; if append = FALSE, only the recoded variables will be returned.

Note

Please note following behaviours of the function:

- the "else"-token should always be the last argument in the rec-string.
- Non-matching values will be set to NA, unless captured by the "else"-token.

- Tagged NA values (see tagged_na) and their value labels will be preserved when copying NA values to the recoded vector with "else=copy".
- Variable label attributes (see, for instance, get_label) are preserved (unless changed via var.label-argument), however, value label attributes are removed (except for "rev", where present value labels will be automatically reversed as well). Use val.labels-argument to add labels for recoded values.
- If x is a data frame, all variables should have the same categories resp. value range (else, see second bullet, NAs are produced).

See Also

set_na for setting NA values, replace_na to replace NA's with specific value, recode_to for re-shifting value ranges and ref_lvl to change the reference level of (numeric) factors.

Examples

```
data(efc)
table(efc$e42dep, useNA = "always")

# replace NA with 5
table(rec(efc$e42dep, rec = "1=1;2=2;3=3;4=4;NA=5"), useNA = "always")

# recode 1 to 2 into 1 and 3 to 4 into 2
table(rec(efc$e42dep, rec = "1,2=1; 3,4=2"), useNA = "always")

# keep value labels. variable label is automatically preserved
library(dplyr)
efc %>%
  select(e42dep) %>%
  rec(rec = "1,2=1; 3,4=2",
      val.labels = c("low dependency", "high dependency")) %>%
  str()

# works with mutate
efc %>%
  select(e42dep, e17age) %>%
  mutate(dependency_rev = rec(e42dep, rec = "rev")) %>%
  head()

# recode 1 to 3 into 4 into 2
table(rec(efc$e42dep, rec = "min:3=1; 4=2"), useNA = "always")

# recode 2 to 1 and all others into 2
table(rec(efc$e42dep, rec = "2=1; else=2"), useNA = "always")

# reverse value order
table(rec(efc$e42dep, rec = "rev"), useNA = "always")

# recode only selected values, copy remaining
table(efc$e15relat)
table(rec(efc$e15relat, rec = "1,2,4=1; else=copy"))
```

```
# recode variables with same category in a data frame
head(efc[, 6:9])
head(rec(efc[, 6:9], rec = "1=10;2=20;3=30;4=40"))

# recode multiple variables and set value labels via recode-syntax
dummy <- rec(
  efc, c160age, e17age,
  rec = "15:30=1 [young]; 31:55=2 [middle]; 56:max=3 [old]",
  append = FALSE
)
frq(dummy)

# recode variables with same value-range
lapply(
  rec(
    efc, c82cop1, c83cop2, c84cop3,
    rec = "1,2=1; NA=9; else=copy",
    append = FALSE
  ),
  table,
  useNA = "always"
)

# recode character vector
dummy <- c("M", "F", "F", "X")
rec(dummy, rec = "M=Male; F=Female; X=Refused")

# recode numeric to character
rec(efc$e42dep, rec = "1=first;2=2nd;3=third;else=hi")

# recode non-numeric factors
data(iris)
table(rec(iris, Species, rec = "setosa=huhu; else=copy", append = FALSE))

# recode floating points
table(rec(
  iris, Sepal.Length, rec = "lo:5=1;5.01:6.5=2;6.501:max=3", append = FALSE
))

# preserve tagged NAs
library(haven)
x <- labelled(c(1:3, tagged_na("a", "c", "z"), 4:1),
              c("Agreement" = 1, "Disagreement" = 4, "First" = tagged_na("c"),
                "Refused" = tagged_na("a"), "Not home" = tagged_na("z")))
# get current value labels
x
# recode 2 into 5; Values of tagged NAs are preserved
rec(x, rec = "2=5;else=copy")
na_tag(rec(x, rec = "2=5;else=copy"))

# use select-helpers from dplyr-package
rec(
```

```
  efc, contains("cop"), c161sex:c175empl,
  rec = "0,1=0; else=1",
  append = FALSE
)
```

---

recode_to *Recode variable categories into new values*

Description

Recodes (or "renumbers") the categories of variables into new category values, beginning with the lowest value specified by `lowest`. Useful if you want to recode dummy variables with 1/2 coding to 0/1 coding, or recoding scales from 1-4 to 0-3 etc.

Usage

```
recode_to(x, ..., lowest = 0, highest = -1, append = TRUE,
  suffix = "_r0")
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
lowest	Indicating the lowest category value for recoding. Default is 0, so the new variable starts with value 0.
highest	If specified and greater than lowest, all category values larger than highest will be set to NA. Default is -1, i.e. this argument is ignored and no NA's will be produced.
append	Logical, if TRUE (the default) and x is a data frame, x including the new variables as additional columns is returned; if FALSE, only the new variables are returned.
suffix	String value, will be appended to variable (column) names of x, if x is a data frame. If x is not a data frame, this argument will be ignored. The default value to suffix column names in a data frame depends on the function call: • recoded variables (rec()) will be suffixed with "_r" • recoded variables (recode_to()) will be suffixed with "_r0" • dichotomized variables (dicho()) will be suffixed with "_d" • grouped variables (split_var()) will be suffixed with "_g" • grouped variables (group_var()) will be suffixed with "_gr" • standardized variables (std()) will be suffixed with "_z" • centered variables (center()) will be suffixed with "_c"

Value

x with recoded category values, where `lowest` indicates the lowest value; If x is a data frame, only the recoded variables will be returned.

Note

Value and variable label attributes are preserved.

See Also

`rec` for general recoding of variables and `set_na` for setting `NA` values.

Examples

```
# recode 1-4 to 0-3
dummy <- sample(1:4, 10, replace = TRUE)
recode_to(dummy)

# recode 3-6 to 0-3
# note that numeric type is returned
dummy <- as.factor(3:6)
recode_to(dummy)

# lowest value starting with 1
dummy <- sample(11:15, 10, replace = TRUE)
recode_to(dummy, lowest = 1)

# lowest value starting with 1, highest with 3
# all others set to NA
dummy <- sample(11:15, 10, replace = TRUE)
recode_to(dummy, lowest = 1, highest = 3)

# recode multiple variables at once
data(efc)
recode_to(efc, c82cop1, c83cop2, c84cop3, append = FALSE)

library(dplyr)
efc %>%
  select(c82cop1, c83cop2, c84cop3) %>%
  mutate(
    c82new = recode_to(c83cop2, lowest = 5),
    c83new = recode_to(c84cop3, lowest = 3)
  ) %>%
  head()
```

rec_pattern *Create recode pattern for 'rec' function*

Description

Convenient function to create a recode pattern for the rec function, which recodes (numeric) vectors into smaller groups.

Usage

```
rec_pattern(from, to, width = 5, other = NULL)
```

Arguments

from	Minimum value that should be recoded.
to	Maximum value that should be recoded.
width	Numeric, indicating the range of each group.
other	String token, indicating how to deal with all other values that have not been captured by the recode pattern. See 'Details' on the else-token in rec.

Value

A list with two values:

pattern string pattern that can be used as rec argument for the rec-function.

labels the associated values labels that can be used with set_labels.

See Also

group_var for recoding variables into smaller groups, and group_labels to create the asssociated value labels.

Examples

```
rp <- rec_pattern(1, 100)
rp

# sample data, inspect age of carers
data(efc)
table(efc$c160age, exclude = NULL)
table(rec(efc$c160age, rec = rp$pattern), exclude = NULL)

# recode carers age into groups of width 5
x <- rec(
  efc$c160age,
  rec = rp$pattern,
  val.labels = rp$labels
)
```

```
# watch result
frq(x)
```

`ref_lvl` *Change reference level of (numeric) factors*

Description

Changes the reference level of numeric factor. See 'Details'.

Usage

```
ref_lvl(x, ..., lvl = NULL)
```

Arguments

`x`	A vector or data frame.
`...`	Optional, unquoted names of variables that should be selected for further processing. Required, if `x` is a data frame (and no vector) and only selected variables from `x` should be processed. You may also use functions like `:` or tidyselect's `select_helpers`. See 'Examples' or package-vignette.
`lvl`	Numeric, the new reference level.

Details

Unlike `relevel`, this function a) only accepts numeric factors and b) changes the reference level by recoding the factor's values using the `rec` function. Hence, all values from lowest up to the reference level indicated by `lvl` are recoded, with `lvl` starting as lowest factor value. See 'Examples'.

Value

`x` with new reference level. If `x` is a data frame, the complete data frame `x` will be returned, where variables specified in `...` will be re-leveled; if `...` is not specified, applies to all variables in the data frame.

See Also

`to_factor` to convert numeric vectors into factors; `rec` to recode variables.

Examples

```
data(efc)
x <- to_factor(efc$e42dep)
str(x)
frq(x)

x <- ref_lvl(x, lvl = 3)
```

```
str(x)
frq(x)

library(dplyr)
dat <- efc %>%
  select(c82cop1, c83cop2, c84cop3) %>%
  to_factor()

frq(dat)
ref_lvl(dat, c82cop1, c83cop2, lvl = 2) %>% frq()
```

`remove_var` *Remove variables from a data frame*

Description

This function removes variables from a data frame, and is intended to use within a pipe-workflow.

Usage

```
remove_var(x, ...)
```

Arguments

`x`	A vector or data frame.
`...`	Character vector with variable names, or unquoted names of variables that should be removed from the data frame. You may also use functions like : or tidyselect's `select_helpers`.

Value

`x`, with variables specified in `...` removed.

Examples

```
mtcars %>% remove_var("disp", "cyl")
mtcars %>% remove_var(c("wt", "vs"))
mtcars %>% remove_var(drat:am)
```

replace_na *Replace NA with specific values*

Description

This function replaces (tagged) NA's of a variable, data frame or list of variables with value.

Usage

```
replace_na(x, ..., value, na.label = NULL, tagged.na = NULL)
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
value	Value that will replace the NA's.
na.label	Optional character vector, used to label the the former NA-value (i.e. adding a labels attribute for value to x).
tagged.na	Optional single character, specifies a tagged_na value that will be replaced by value. Herewith it is possible to replace only specific NA values of x.

Details

While regular NA values can only be *completely* replaced with a single value, tagged_na allows to differentiate between different qualitative values of NAs. Tagged NAs work exactly like regular R missing values except that they store one additional byte of information: a tag, which is usually a letter ("a" to "z") or character number ("0" to "9"). Therewith it is possible to replace only specific NA values, while other NA values are preserved.

Value

x, where NA's are replaced with value. If x is a data frame, the complete data frame x will be returned, with replaced NA's for variables specified in ...; if ... is not specified, applies to all variables in the data frame.

Note

Value and variable label attributes are preserved.

See Also

set_na for setting NA values, rec for general recoding of variables and recode_to for re-shifting value ranges.

Examples

```
library(sjlabelled)
data(efc)
table(efc$e42dep, useNA = "always")
table(replace_na(efc$e42dep, value = 99), useNA = "always")

# the original labels
get_labels(replace_na(efc$e42dep, value = 99))
# NA becomes "99", and is labelled as "former NA"
get_labels(replace_na(efc$e42dep, value = 99, na.label = "former NA"),
           include.values = "p")

dummy <- data.frame(
  v1 = efc$c82cop1,
  v2 = efc$c83cop2,
  v3 = efc$c84cop3
)
# show original distribution
lapply(dummy, table, useNA = "always")
# show variables, NA's replaced with 99
lapply(replace_na(dummy, v2, v3, value = 99), table, useNA = "always")

library(haven)
x <- labelled(c(1:3, tagged_na("a", "c", "z"), 4:1),
              c("Agreement" = 1, "Disagreement" = 4, "First" = tagged_na("c"),
                "Refused" = tagged_na("a"), "Not home" = tagged_na("z")))
# get current NA values
x
get_na(x)

# replace only the NA, which is tagged as NA(c)
replace_na(x, value = 2, tagged.na = "c")
get_na(replace_na(x, value = 2, tagged.na = "c"))

table(x)
table(replace_na(x, value = 2, tagged.na = "c"))

# tagged NA also works for non-labelled class
# init vector
x <- c(1, 2, 3, 4)
# set values 2 and 3 as NA, will automatically become
# tagged NAs by 'set_na()'.
x <- set_na(x, na = c(2, 3))
# see result
x
# now replace only NA tagged with 2 with value 5
replace_na(x, value = 5, tagged.na = "2")
```

`rotate_df` *Rotate a data frame*

Description

This function rotates a data frame, i.e. columns become rows and vice versa.

Usage

```
rotate_df(x, rn = NULL, cn = FALSE)
```

Arguments

x	A data frame.
rn	Character vector (optional). If not NULL, the data frame's rownames will be added as (first) column to the output, with rn being the name of this column.
cn	Logical (optional), if TRUE, the values of the first column in x will be used as column names in the rotated data frame.

Value

A (rotated) data frame.

Examples

```
x <- mtcars[1:3, 1:4]
rotate_df(x)
rotate_df(x, rn = "property")

# use values in 1. column as column name
library(tibble)
x <- tibble::rownames_to_column(x)
rotate_df(x)
rotate_df(x, cn = TRUE)
rotate_df(x, rn = "property", cn = TRUE)

# also works on list-results
library(purrr)

dat <- mtcars[1:3, 1:4]
tmp <- purrr::map(dat, function(x) {
  sdev <- stats::sd(x, na.rm = TRUE)
  ulsdev <- mean(x, na.rm = TRUE) + c(-sdev, sdev)
  names(ulsdev) <- c("lower_sd", "upper_sd")
  ulsdev
})
tmp
as.data.frame(tmp)
rotate_df(tmp)

tmp <- purrr::map_df(dat, function(x) {
  sdev <- stats::sd(x, na.rm = TRUE)
  ulsdev <- mean(x, na.rm = TRUE) + c(-sdev, sdev)
  names(ulsdev) <- c("lower_sd", "upper_sd")
  ulsdev
```

```
})
tmp
rotate_df(tmp)
```

---

row_count *Count row or column indices*

Description

row_count() mimics base R's rowSums(), with sums for a specific value indicated by count. Hence, it is equivalent to rowSums(x == count, na.rm = TRUE). However, this function is designed to work nicely within a pipe-workflow and allows select-helpers for selecting variables and the return value is always a tibble (with one variable).

col_count() does the same for columns. The return value is a data frame with one row (the column counts) and the same number of columns as x.

Usage

```
row_count(x, ..., count, var = "rowcount", append = TRUE)

col_count(x, ..., count, var = "colcount", append = TRUE)
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
count	The value for which the row or column sum should be computed. May be a numeric value, a character string (for factors or character vectors), NA, Inf or NULL to count missing or infinite values, or null-values.
var	Name of new the variable with the row or column counts.
append	Logical, if TRUE (the default) and x is a data frame, x including the new variables as additional columns is returned; if FALSE, only the new variables are returned.

Value

For row_count(), a tibble with one variable: the sum of count appearing in each row of x; for col_count(), a tibble with one row and the same number of variables as in x: each variable holds the sum of count appearing in each variable of x. If append = TRUE, x including this variable will be returned.

Examples

```
library(dplyr)
library(tibble)
dat <- tribble(
  ~c1, ~c2, ~c3, ~c4,
    1,   3,   1,   1,
    2,   2,   1,   1,
    3,   1,   2,   3,
    1,   2,   1,   2,
    3,  NA,   3,   1,
   NA,   3,  NA,   2
)

row_count(dat, count = 1, append = FALSE)
row_count(dat, count = NA, append = FALSE)
row_count(dat, c1:c3, count = 2, append = TRUE)

col_count(dat, count = 1, append = FALSE)
col_count(dat, count = NA, append = FALSE)
col_count(dat, c1:c3, count = 2, append = TRUE)
```

`row_sums` *Row sums and means for data frames*

Description

`row_sums()` simply wraps `rowSums`, while `row_means()` simply wraps `mean_n`, however, the argument-structure of both functions is designed to work nicely within a pipe-workflow and allows select-helpers for selecting variables, the default for `na.rm` is `TRUE`, and the return value is always a tibble (with one variable).

Usage

```
row_sums(x, ..., na.rm = TRUE, var = "rowsums", append = TRUE)

row_means(x, ..., n, var = "rowmeans", append = TRUE)
```

Arguments

`x`	A vector or data frame.
`...`	Optional, unquoted names of variables that should be selected for further processing. Required, if `x` is a data frame (and no vector) and only selected variables from `x` should be processed. You may also use functions like `:` or tidyselect's `select_helpers`. See 'Examples' or package-vignette.
`na.rm`	Logical, `TRUE` if missing values should be omitted from the calculations.
`var`	Name of new the variable with the row sums or means.

append — Logical, if TRUE (the default) and x is a data frame, x including the new variables as additional columns is returned; if FALSE, only the new variables are returned.

n — May either be

- a numeric value that indicates the amount of valid values per row to calculate the row mean;
- or a value between 0 and 1, indicating a proportion of valid values per row to calculate the row mean (see 'Details').

If a row's sum of valid values is less than n, NA will be returned as row mean value.

Details

For n, must be a numeric value from 0 to ncol(x). If a *row* in x has at least n non-missing values, the row mean is returned. If n is a non-integer value from 0 to 1, n is considered to indicate the proportion of necessary non-missing values per row. E.g., if n = .75, a row must have at least ncol(x) * n non-missing values for the row mean to be calculated. See 'Examples'.

Value

For row_sums(), a tibble with a new variable: the row sums from x; for row_means(), a tibble with a new variable: the row means from x. If append = FALSE, only the new variable with row sums resp. row means is returned.

Examples

```
data(efc)
efc %>% row_sums(c82cop1:c90cop9, append = FALSE)

library(dplyr)
row_sums(efc, contains("cop"), append = FALSE)

dat <- data.frame(
  c1 = c(1,2,NA,4),
  c2 = c(NA,2,NA,5),
  c3 = c(NA,4,NA,NA),
  c4 = c(2,3,7,8),
  c5 = c(1,7,5,3)
)
dat

row_means(dat, n = 4)
row_means(dat, c1:c4, n = 4)
# at least 40% non-missing
row_means(dat, c1:c4, n = .4)

# create sum-score of COPE-Index, and append to data
efc %>%
  select(c82cop1:c90cop9) %>%
  row_sums()
```

set_na *Replace specific values in vector with NA*

Description

This function replaces specific values of variables with NA.

Usage

```
set_na(x, ..., na, drop.levels = TRUE, as.tag = FALSE)
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
na	Numeric vector with values that should be replaced with NA values, or a character vector if values of factors or character vectors should be replaced. For labelled vectors, may also be the name of a value label. In this case, the associated values for the value labels in each vector will be replaced with NA (see 'Examples').
drop.levels	Logical, if TRUE, factor levels of values that have been replaced with NA are dropped. See 'Examples'.
as.tag	Logical, if TRUE, values in x will be replaced by tagged_na, else by usual NA values. Use a named vector to assign the value label to the tagged NA value (see 'Examples').

Details

set_na() converts all values defined in na with a related NA or tagged NA value (see tagged_na). Tagged NAs work exactly like regular R missing values except that they store one additional byte of information: a tag, which is usually a letter ("a" to "z") or character number ("0" to "9").

Furthermore, see also 'Details' in get_na.

Value

x, with all elements of na being replaced by NA. If x is a data frame, the complete data frame x will be returned, with NA's set for variables specified in ...; if ... is not specified, applies to all variables in the data frame.

Note

Labels from values that are replaced with NA and no longer used will be removed from x, however, other value and variable label attributes are preserved. For more details on labelled data, see vignette Labelled Data and the sjlabelled-Package.

See Also

replace_na to replace NA's with specific values, rec for general recoding of variables and recode_to for re-shifting value ranges. See get_na to get values of missing values in labelled vectors.

Examples

```
# create random variable
dummy <- sample(1:8, 100, replace = TRUE)
# show value distribution
table(dummy)
# set value 1 and 8 as missings
dummy <- set_na(dummy, na = c(1, 8))
# show value distribution, including missings
table(dummy, useNA = "always")

# add named vector as further missing value
set_na(dummy, na = c("Refused" = 5), as.tag = TRUE)
# see different missing types
library(haven)
library(sjlabelled)
print_tagged_na(set_na(dummy, na = c("Refused" = 5), as.tag = TRUE))


# create sample data frame
dummy <- data.frame(var1 = sample(1:8, 100, replace = TRUE),
                    var2 = sample(1:10, 100, replace = TRUE),
                    var3 = sample(1:6, 100, replace = TRUE))
# set value 2 and 4 as missings
dummy %>% set_na(na = c(2, 4)) %>% head()
dummy %>% set_na(na = c(2, 4), as.tag = TRUE) %>% get_na()
dummy %>% set_na(na = c(2, 4), as.tag = TRUE) %>% get_values()

data(efc)
dummy <- data.frame(
  var1 = efc$c82cop1,
  var2 = efc$c83cop2,
  var3 = efc$c84cop3
)
# check original distribution of categories
lapply(dummy, table, useNA = "always")
# set 3 to NA for two variables
lapply(set_na(dummy, var1, var3, na = 3), table, useNA = "always")

# drop unused factor levels when being set to NA
x <- factor(c("a", "b", "c"))
x
set_na(x, na = "b", as.tag = TRUE)
set_na(x, na = "b", drop.levels = FALSE, as.tag = TRUE)

# set_na() can also remove a missing by defining the value label
# of the value that should be replaced with NA. This is in particular
# helpful if a certain category should be set as NA, however, this category
```

```
# is assigned with different values accross variables
x1 <- sample(1:4, 20, replace = TRUE)
x2 <- sample(1:7, 20, replace = TRUE)
x1 <- set_labels(x1, labels = c("Refused" = 3, "No answer" = 4))
x2 <- set_labels(x2, labels = c("Refused" = 6, "No answer" = 7))

tmp <- data.frame(x1, x2)
get_labels(tmp)
get_labels(set_na(tmp, na = "No answer"))
get_labels(set_na(tmp, na = c("Refused", "No answer")))

# show values
tmp
set_na(tmp, na = c("Refused", "No answer"))
```

`shorten_string` *Shorten character strings*

Description

This function shortens strings that are longer than `max.length` chars.

Usage

```
shorten_string(s, max.length = NULL, abbr = "...")
```

Arguments

`s`	A string.
`max.length`	Maximum length of chars for the string.
`abbr`	String that will be used as suffix, if `s` was shortened.

Value

A shortened string.

Examples

```
s <- "This can be considered as very long string!"

# string is shorter than max.length, so returned as is
shorten_string(s, 60)

# string is shortened to as many words that result in
# a string of maximum 20 chars
shorten_string(s, 20)
```

```
# string including "considered" is exactly of length 22 chars
shorten_string(s, 22)
```

---

split_var *Split numeric variables into smaller groups*

Description

Recode numeric variables into equal sized groups, i.e. a variable is cut into a smaller number of groups at specific cut points.

Usage

```
split_var(x, ..., n, as.num = FALSE, val.labels = NULL, var.label = NULL,
  inclusive = FALSE, append = TRUE, suffix = "_g")
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
n	The new number of groups that x should be split into.
as.num	Logical, if TRUE, return value will be numeric, not a factor.
val.labels	Optional character vector, to set value label attributes of recoded variable (see vignette Labelled Data and the sjlabelled-Package). If NULL (default), no value labels will be set. Value labels can also be directly defined in the rec-syntax, see 'Details'.
var.label	Optional string, to set variable label attribute for the returned variable (see vignette Labelled Data and the sjlabelled-Package). If NULL (default), variable label attribute of x will be used (if present). If empty, variable label attributes will be removed.
inclusive	Logical; if TRUE, cut point value are included in the preceeding group. This may be necessary if cutting a vector into groups does not define proper ("equal sized") group sizes. See 'Note' and 'Examples'.
append	Logical, if TRUE (the default) and x is a data frame, x including the new variables as additional columns is returned; if FALSE, only the new variables are returned.
suffix	String value, will be appended to variable (column) names of x, if x is a data frame. If x is not a data frame, this argument will be ignored. The default value to suffix column names in a data frame depends on the function call: • recoded variables (rec()) will be suffixed with "_r" • recoded variables (recode_to()) will be suffixed with "_r0"

- dichotomized variables (dicho()) will be suffixed with "_d"
- grouped variables (split_var()) will be suffixed with "_g"
- grouped variables (group_var()) will be suffixed with "_gr"
- standardized variables (std()) will be suffixed with "_z"
- centered variables (center()) will be suffixed with "_c"

Details

split_var() splits a variable into equal sized groups, where the amount of groups depends on the groupcount-argument. Thus, this functions cuts a variable into groups at the specified quantiles.

By contrast, group_var recodes a variable into groups, where groups have the same value range (e.g., from 1-5, 6-10, 11-15 etc.).

split_var() also works on grouped data frames (see group_by). In this case, splitting is applied to the subsets of variables in x. See 'Examples'.

Value

A grouped variable with equal sized groups. If x is a data frame, for append = TRUE, x including the grouped variables as new columns is returned; if append = FALSE, only the grouped variables will be returned.

Note

In case a vector has only few number of unique values, splitting into equal sized groups may fail. In this case, use the inclusive-argument to shift a value at the cut point into the lower, preceeding group to get equal sized groups. See 'Examples'.

See Also

group_var to group variables into equal ranged groups, or rec to recode variables.

Examples

```
data(efc)
# non-grouped
table(efc$neg_c_7)

# split into 3 groups
table(split_var(efc$neg_c_7, n = 3))

# split multiple variables into 3 groups
split_var(efc, neg_c_7, pos_v_4, e17age, n = 3, append = FALSE)
frq(split_var(efc, neg_c_7, pos_v_4, e17age, n = 3, append = FALSE))

# original
table(efc$e42dep)

# two groups, non-inclusive cut-point
```

```
# vector split leads to unequal group sizes
table(split_var(efc$e42dep, n = 2))

# two groups, inclusive cut-point
# group sizes are equal
table(split_var(efc$e42dep, n = 2, inclusive = TRUE))

# Unlike dplyr's ntile(), split_var() never splits a value
# into two different categories, i.e. you always get a clean
# separation of original categories
library(dplyr)

x <- dplyr::ntile(efc$neg_c_7, n = 3)
table(efc$neg_c_7, x)

x <- split_var(efc$neg_c_7, n = 3)
table(efc$neg_c_7, x)

# works also with gouped data frames
mtcars %>%
  split_var(disp, n = 3, append = FALSE) %>%
  table()

mtcars %>%
  group_by(cyl) %>%
  split_var(disp, n = 3, append = FALSE) %>%
  table()
```

`spread_coef` *Spread model coefficients of list-variables into columns*

Description

This function extracts coefficients (and standard error and p-values) of fitted model objects from (nested) data frames, which are saved in a list-variable, and spreads the coefficients into new columnns.

Usage

```
spread_coef(data, model.column, model.term, se, p.val, append = TRUE, ...)
```

Arguments

`data`	A (nested) data frame with a list-variable that contains fitted model objects (see 'Details').
`model.column`	Name or index of the list-variable that contains the fitted model objects.
`model.term`	Optional, name of a model term. If specified, only this model term (including p-value) will be extracted from each model and added as new column.

se	Logical, if TRUE, standard errors for estimates will also be extracted.
p.val	Logical, if TRUE, p-values for estimates will also be extracted.
append	Logical, if TRUE (default), this function returns data with new columns for the model coefficients; else, a new data frame with model coefficients only are returned.
...	Other arguments passed down to the tidy-function.

Details

This function requires a (nested) data frame (e.g. created by the nest-function of the **tidyr**-package), where several fitted models are saved in a list-variable (see 'Examples'). Since nested data frames with fitted models stored as list-variable are typically fit with an identical formula, all models have the same dependent and independent variables and only differ in their subsets of data. The function then extracts all coefficients from each model and saves each estimate in a new column. The result is a data frame, where each *row* is a model with each model's coefficients in an own *column*.

Value

A data frame with columns for each coefficient of the models that are stored in the list-variable of data; or, if model.term is given, a data frame with the term's estimate. If se = TRUE or p.val = TRUE, the returned data frame also contains columns for the coefficients' standard error and p-value. If append = TRUE, the columns are appended to data, i.e. data is also returned.

Examples

```
library(dplyr)
library(tidyr)
library(purrr)
data(efc)

# create nested data frame, grouped by dependency (e42dep)
# and fit linear model for each group. These models are
# stored in the list variable "models".
model.data <- efc %>%
  filter(!is.na(e42dep)) %>%
  group_by(e42dep) %>%
  nest() %>%
  mutate(
    models = map(data, ~lm(neg_c_7 ~ c12hour + c172code, data = .x))
  )

# spread coefficients, so we can easily access and compare the
# coefficients over all models. arguments `se` and `p.val` default
# to `FALSE`, when `model.term` is not specified
spread_coef(model.data, models)
spread_coef(model.data, models, se = TRUE)

# select only specific model term. `se` and `p.val` default to `TRUE`
spread_coef(model.data, models, c12hour)
```

```
# spread_coef can be used directly within a pipe-chain
efc %>%
  filter(!is.na(e42dep)) %>%
  group_by(e42dep) %>%
  nest() %>%
  mutate(
    models = map(data, ~lm(neg_c_7 ~ c12hour + c172code, data = .x))
  ) %>%
  spread_coef(models)

# spread_coef() makes it easy to generate bootstrapped
# confidence intervals, using the 'bootstrap()' and 'boot_ci()'
# functions from the 'sjstats' package, which creates nested
# data frames of bootstrap replicates
library(sjstats)
efc %>%
  # generate bootstrap replicates
  bootstrap(100) %>%
  # apply lm to all bootstrapped data sets
  mutate(
    models = map(strap, ~lm(neg_c_7 ~ e42dep + c161sex + c172code, data = .x))
  ) %>%
  # spread model coefficient for all 100 models
  spread_coef(models, se = FALSE, p.val = FALSE) %>%
  # compute the CI for all bootstrapped model coefficients
  boot_ci(e42dep, c161sex, c172code)
```

`std` *Standardize and center variables*

Description

`std()` computes a z-transformation (standardized and centered) on the input. `center()` centers the input.

Usage

```
std(x, ..., robust = c("sd", "gmd", "mad"), include.fac = FALSE,
  append = TRUE, suffix = "_z")

center(x, ..., include.fac = FALSE, append = TRUE, suffix = "_c")
```

Arguments

x	A vector or data frame.

`...` Optional, unquoted names of variables that should be selected for further processing. Required, if `x` is a data frame (and no vector) and only selected variables from `x` should be processed. You may also use functions like `:` or tidyselect's `select_helpers`. See 'Examples' or package-vignette.

`robust` Character vector, indicating the method applied when standardizing variables with `std()`. By default, standardization is achieved by dividing the centered variables by their standard deviation (`robust = "sd"`). However, for skewed distributions, the median absolute deviation (MAD, `robust = "mad"`) or Gini's mean difference (`robust = "gmd"`) might be more robust measures of dispersion. For the latter option, **sjstats** needs to be installed.

`include.fac` Logical, if `TRUE`, factors will be converted to numeric vectors and also standardized or centered.

`append` Logical, if `TRUE` (the default) and `x` is a data frame, `x` including the new variables as additional columns is returned; if `FALSE`, only the new variables are returned.

`suffix` String value, will be appended to variable (column) names of `x`, if `x` is a data frame. If `x` is not a data frame, this argument will be ignored. The default value to suffix column names in a data frame depends on the function call:

- recoded variables (`rec()`) will be suffixed with `"_r"`
- recoded variables (`recode_to()`) will be suffixed with `"_r0"`
- dichotomized variables (`dicho()`) will be suffixed with `"_d"`
- grouped variables (`split_var()`) will be suffixed with `"_g"`
- grouped variables (`group_var()`) will be suffixed with `"_gr"`
- standardized variables (`std()`) will be suffixed with `"_z"`
- centered variables (`center()`) will be suffixed with `"_c"`

Details

`std()` and `center()` also work on grouped data frames (see `group_by`). In this case, standardization or centering is applied to the subsets of variables in `x`. See 'Examples'.

Value

A vector with standardized or centered variables. If `x` is a data frame, only the transformed variables will be returned.

Note

`std()` and `center()` only return a vector, if `x` is a vector. If `x` is a data frame and only one variable is specified in the `...`-ellipses argument, both functions do return a data frame (see 'Examples').

Examples

```
data(efc)
std(efc$c160age) %>% head()
std(efc, e17age, c160age, append = FALSE) %>% head()

center(efc$c160age) %>% head()
```

```
center(efc, e17age, c160age, append = FALSE) %>% head()

# NOTE!
std(efc$e17age) # returns a vector
std(efc, e17age) # returns a tibble

# works with mutate()
library(dplyr)
efc %>%
  select(e17age, neg_c_7) %>%
  mutate(age_std = std(e17age), burden = center(neg_c_7)) %>%
  head()

# works also with grouped data frames
mtcars %>% std(disp)

mtcars %>%
  group_by(cyl) %>%
  std(disp)

data(iris)
# also standardize factors
std(iris, include.fac = TRUE, append = FALSE)
# don't standardize factors
std(iris, include.fac = FALSE, append = FALSE)
```

`str_contains` *Check if string contains pattern*

Description

This functions checks whether a string or character vector `x` contains the string `pattern`. By default, this function is case sensitive.

Usage

```
str_contains(x, pattern, ignore.case = FALSE, logic = NULL,
  switch = FALSE)
```

Arguments

`x`	Character string where matches are sought. May also be a character vector of length > 1 (see 'Examples').
`pattern`	Character string to be matched in `x`. May also be a character vector of length > 1 (see 'Examples').
`ignore.case`	Logical, whether matching should be case sensitive or not.
`logic`	Indicates whether a logical combination of multiple search pattern should be made.

- Use "or", "OR" or "|" for a logical or-combination, i.e. at least one element of pattern is in x.
- Use "and", "AND" or "&" for a logical AND-combination, i.e. all elements of pattern are in x.
- Use "not", "NOT" or "!" for a logical NOT-combination, i.e. no element of pattern is in x.
- By default, logic = NULL, which means that TRUE or FALSE is returned for each element of pattern separately.

switch Logical, if TRUE, x will be sought in each element of pattern. If switch = TRUE, x needs to be of length 1.

Details

This function iterates all elements in pattern and looks for each of these elements if it is found in *any* element of x, i.e. which elements of pattern are found in the vector x.

Technically, it iterates pattern and calls grep(x, pattern[i], fixed = TRUE) for each element of pattern. If switch = TRUE, it iterates pattern and calls grep(pattern[i], x, fixed = TRUE) for each element of pattern. Hence, in the latter case (if switch = TRUE), x must be of length 1.

Value

TRUE if x contains pattern.

Examples

```
str_contains("hello", "hel")
str_contains("hello", "hal")

str_contains("hello", "Hel")
str_contains("hello", "Hel", ignore.case = TRUE)

# which patterns are in "abc"?
str_contains("abc", c("a", "b", "e"))

# is pattern in any element of 'x'?
str_contains(c("def", "abc", "xyz"), "abc")
# is "abcde" in any element of 'x'?
str_contains(c("def", "abc", "xyz"), "abcde") # no...
# is "abc" in any of pattern?
str_contains("abc", c("defg", "abcde", "xyz12"), switch = TRUE)

str_contains(c("def", "abcde", "xyz"), c("abc", "123"))

# any pattern in "abc"?
str_contains("abc", c("a", "b", "e"), logic = "or")

# all patterns in "abc"?
str_contains("abc", c("a", "b", "e"), logic = "and")
str_contains("abc", c("a", "b"), logic = "and")
```

```
# no patterns in "abc"?
str_contains("abc", c("a", "b", "e"), logic = "not")
str_contains("abc", c("d", "e", "f"), logic = "not")
```

`str_pos` *Find partial matching and close distance elements in strings*

Description

This function finds the element indices of partial matching or similar strings in a character vector. Can be used to find exact or slightly mistyped elements in a string vector.

Usage

```
str_pos(search.string, find.term, maxdist = 2, part.dist.match = 0,
  show.pbar = FALSE)
```

Arguments

`search.string`	Character vector with string elements.
`find.term`	String that should be matched against the elements of `search.string`.
`maxdist`	Maximum distance between two string elements, which is allowed to treat them as similar or equal. Smaller values mean less tolerance in matching.
`part.dist.match`	Activates similar matching (close distance strings) for parts (substrings) of the `search.string`. Following values are accepted: • 0 for no partial distance matching • 1 for one-step matching, which means, only substrings of same length as `find.term` are extracted from `search.string` matching • 2 for two-step matching, which means, substrings of same length as `find.term` as well as strings with a slightly wider range are extracted from `search.string` matching Default value is 0. See 'Details' for more information.
`show.pbar`	Logical; f `TRUE`, the progress bar is displayed when computing the distance matrix. Default in `FALSE`, hence the bar is hidden.

Details

For `part.dist.match = 1`, a substring of `length(find.term)` is extracted from `search.string`, starting at position 0 in `search.string` until the end of `search.string` is reached. Each substring is matched against `find.term`, and results with a maximum distance of `maxdist` are considered as "matching". If `part.dist.match = 2`, the range of the extracted substring is increased by 2, i.e. the extracted substring is two chars longer and so on.

Value

A numeric vector with index position of elements in `search.string` that partially match or are similar to `find.term`. Returns `-1` if no match was found.

Note

This function does *not* return the position of a matching string *inside* another string, but the element's index of the `search.string` vector, where a (partial) match with `find.term` was found. Thus, searching for "abc" in a string "this is abc" will not return 9 (the start position of the substring), but 1 (the element index, which is always 1 if `search.string` only has one element).

See Also

`group_str`

Examples

```
## Not run:
string <- c("Hello", "Helo", "Hole", "Apple", "Ape", "New", "Old", "System", "Systemic")
str_pos(string, "hel")   # partial match
str_pos(string, "stem")  # partial match
str_pos(string, "R")     # no match
str_pos(string, "saste") # similarity to "System"

# finds two indices, because partial matching now
# also applies to "Systemic"
str_pos(string,
        "sytsme",
        part.dist.match = 1)

# finds nothing
str_pos("We are Sex Pistols!", "postils")
# finds partial matching of similarity
str_pos("We are Sex Pistols!", "postils", part.dist.match = 1)
## End(Not run)
```

`str_start` *Find start and end index of pattern in string*

Description

`str_start()` finds the beginning position of `pattern` in each element of `x`, while `str_end()` finds the stopping position of `pattern` in each element of `x`.

Usage

```
str_start(x, pattern, ignore.case = TRUE)

str_end(x, pattern, ignore.case = TRUE)
```

Arguments

x	A character vector.
pattern	Character string to be matched in x. pattern might also be a regular-expression object, as returned by regex, or any of **stringr**'s supported modifiers.
ignore.case	Logical, whether matching should be case sensitive or not.

Value

A numeric vector with index of start/end position(s) of pattern found in x, or an empty vector, if pattern was not found in x.

Examples

```
path <- "this/is/my/fileofinterest.csv"
str_start(path, "/")

path <- "this//is//my//fileofinterest.csv"
str_start(path, "//")
str_end(path, "//")

x <- c("my_friend_likes me", "your_friend likes_you")
str_start(x, "_")

# pattern "likes" starts at position 11 in first, and
# position 13 in second string
str_start(x, "likes")

# pattern "likes" ends at position 15 in first, and
# position 17 in second string
str_end(x, "likes")

x <- c("I like to move it, move it", "You like to move it")
str_start(x, "move")
str_end(x, "move")
```

---

to_character *Convert variable into character vector and replace values with associated value labels*

Description

This function converts (replaces) variable values (also of factors or character vectors) with their associated value labels and returns them as character vector. This is just a convenient wrapper for as.character(to_label(x)).

Usage

```
to_character(x, ..., add.non.labelled = FALSE, prefix = FALSE,
  var.label = NULL, drop.na = TRUE, drop.levels = FALSE)
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
add.non.labelled	Logical, if TRUE, values without associated value label will also be converted to labels (as is). See 'Examples'.
prefix	Logical, if TRUE, the value labels used as factor levels or character values will be prefixed with their associated values. See 'Examples'.
var.label	Optional string, to set variable label attribute for the returned variable (see vignette Labelled Data and the sjlabelled-Package). If NULL (default), variable label attribute of x will be used (if present). If empty, variable label attributes will be removed.
drop.na	Logical, if TRUE, tagged NA values with value labels will be converted to regular NA's. Else, tagged NA values will be replaced with their value labels. See 'Examples' and get_na.
drop.levels	Logical, if TRUE, unused factor levels will be dropped (i.e. droplevels will be applied before returning the result).

Value

A character vector with the associated value labels as values. If x is a data frame, the complete data frame x will be returned, where variables specified in ... are coerced to character variables; if ... is not specified, applies to all variables in the data frame.

Note

Value labels will be removed when converting variables to factors, variable labels, however, are preserved.

This function is kept for backwards-compatibility. It is preferred to use as_character.

Examples

```
library(sjlabelled)
data(efc)
print(get_labels(efc)['c161sex'])
head(efc$c161sex)
head(to_character(efc$c161sex))

print(get_labels(efc)['e42dep'])
```

```
table(efc$e42dep)
table(to_character(efc$e42dep))

head(efc$e42dep)
head(to_character(efc$e42dep))

# numeric values w/o value labels will also be converted into character
str(efc$e17age)
str(to_character(efc$e17age))


# factor with non-numeric levels, non-prefixed and prefixed
x <- factor(c("a", "b", "c"))
x <- set_labels(x, labels = c("ape", "bear", "cat"))

to_character(x, prefix = FALSE)
to_character(x, prefix = TRUE)


# create vector
x <- c(1, 2, 3, 2, 4, NA)
# add less labels than values
x <- set_labels(x,
                labels = c("yes", "maybe", "no"),
                force.labels = FALSE,
                force.values = FALSE)
# convert to character w/o non-labelled values
to_character(x)
# convert to character, including non-labelled values
to_character(x, add.non.labelled = TRUE)


# create labelled integer, with missing flag
library(haven)
x <- labelled(c(1:3, tagged_na("a", "c", "z"), 4:1, 2:3),
              c("Agreement" = 1, "Disagreement" = 4, "First" = tagged_na("c"),
                "Refused" = tagged_na("a"), "Not home" = tagged_na("z")))
# to character, with missing labels
to_character(x, drop.na = FALSE)
# to character, missings removed
to_character(x, drop.na = TRUE)
# keep missings, and use non-labelled values as well
to_character(x, add.non.labelled = TRUE, drop.na = FALSE)


# easily coerce specific variables in a data frame to character
# and keep other variables, with their class preserved
to_character(efc, e42dep, e16sex, c172code)
```

---

to_dummy *Split (categorical) vectors into dummy variables*

Description

This function splits categorical or numeric vectors with more than two categories into 0/1-coded dummy variables.

Usage

```
to_dummy(x, ..., var.name = "name", suffix = c("numeric", "label"))
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
var.name	Indicates how the new dummy variables are named. Use "name" to use the variable name or any other string that will be used as is. Only applies, if x is a vector. See 'Examples'.
suffix	Indicates which suffix will be added to each dummy variable. Use "numeric" to number dummy variables, e.g. *x_1*, *x_2*, *x_3* etc. Use "label" to add value label, e.g. *x_low*, *x_mid*, *x_high*. May be abbreviated.

Value

A data frame with dummy variables for each category of x. The dummy coded variables are of type atomic.

Note

NA values will be copied from x, so each dummy variable has the same amount of NA's at the same position as x.

Examples

```
data(efc)
head(to_dummy(efc$e42dep))

# add value label as suffix to new variable name
head(to_dummy(efc$e42dep, suffix = "label"))

# use "dummy" as new variable name
head(to_dummy(efc$e42dep, var.name = "dummy"))

# create multiple dummies, append to data frame
to_dummy(efc, c172code, e42dep)

# pipe-workflow
library(dplyr)
efc %>%
```

```
  select(e42dep, e16sex, c172code) %>%
  to_dummy()
```

`to_factor` *Convert variable into factor and keep value labels*

Description

This function converts a variable into a factor, but preserves variable and value label attributes. See 'Examples'.

Usage

```
to_factor(x, ..., add.non.labelled = FALSE, ref.lvl = NULL)
```

Arguments

`x`	A vector or data frame.
`...`	Optional, unquoted names of variables that should be selected for further processing. Required, if `x` is a data frame (and no vector) and only selected variables from `x` should be processed. You may also use functions like : or tidyselect's `select_helpers`. See 'Examples' or package-vignette.
`add.non.labelled`	Logical, if `TRUE`, non-labelled values also get value labels.
`ref.lvl`	Numeric, specifies the reference level for the new factor. Use this parameter if a different factor level than the lowest value should be used as reference level. If `NULL`, lowest value will become the reference level. See `ref_lvl` for details.

Details

`to_factor` converts numeric values into a factor with numeric levels. `to_label`, however, converts a vector into a factor and uses value labels as factor levels.

Value

A factor, including variable and value labels. If `x` is a data frame, the complete data frame `x` will be returned, where variables specified in `...` are coerced to factors (including variable and value labels); if `...` is not specified, applies to all variables in the data frame.

Note

This function is intended for use with vectors that have value and variable label attributes. Unlike `as.factor`, `to_factor` converts a variable into a factor and preserves the value and variable label attributes.

Adding label attributes is automatically done by importing data sets with one of the read_*-functions, like read_spss. Else, value and variable labels can be manually added to vectors with set_labels and set_label.

This function is kept for backwards-compatibility. It is preferred to use as_factor.

See Also

to_value to convert a factor into a numeric vector and to_label to convert a vector into a factor with labelled factor levels.

Examples

```
library(sjlabelled)
data(efc)
# normal factor conversion, loses value attributes
x <- as.factor(efc$e42dep)
frq(x)

# factor conversion, which keeps value attributes
x <- to_factor(efc$e42dep)
frq(x)

# create parially labelled vector
x <- set_labels(efc$e42dep,
                labels = c(`1` = "independent", `4` = "severe dependency",
                           `9` = "missing value"))

# only copy existing value labels
to_factor(x)
get_labels(to_factor(x), include.values = "p")

# also add labels to non-labelled values
to_factor(x, add.non.labelled = TRUE)
get_labels(to_factor(x, add.non.labelled = TRUE), include.values = "p")


# Convert to factor, using different reference level
x <- to_factor(efc$e42dep)
str(x)
table(x)

x <- to_factor(efc$e42dep, ref.lvl = 3)
str(x)
table(x)


# easily coerce specific variables in a data frame to factor
# and keep other variables, with their class preserved
to_factor(efc, e42dep, e16sex, c172code)

# use select-helpers from dplyr-package
```

```
library(dplyr)
to_factor(efc, contains("cop"), c161sex:c175empl)
```

`to_label` *Convert variable into factor with associated value labels*

Description

This function converts (replaces) values of a variable (also of factors or character vectors) with their associated value labels. Might be helpful for factor variables. For instance, if you have a Gender variable with 0/1 value, and associated labels are male/female, this function would convert all 0 to male and all 1 to female and returns the new variable as factor.

Usage

```
to_label(x, ..., add.non.labelled = FALSE, prefix = FALSE,
  var.label = NULL, drop.na = TRUE, drop.levels = FALSE)
```

Arguments

`x`	A vector or data frame.
`...`	Optional, unquoted names of variables that should be selected for further processing. Required, if `x` is a data frame (and no vector) and only selected variables from `x` should be processed. You may also use functions like : or tidyselect's `select_helpers`. See 'Examples' or package-vignette.
`add.non.labelled`	Logical, if `TRUE`, values without associated value label will also be converted to labels (as is). See 'Examples'.
`prefix`	Logical, if `TRUE`, the value labels used as factor levels or character values will be prefixed with their associated values. See 'Examples'.
`var.label`	Optional string, to set variable label attribute for the returned variable (see vignette Labelled Data and the sjlabelled-Package). If `NULL` (default), variable label attribute of `x` will be used (if present). If empty, variable label attributes will be removed.
`drop.na`	Logical, if `TRUE`, tagged `NA` values with value labels will be converted to regular NA's. Else, tagged `NA` values will be replaced with their value labels. See 'Examples' and `get_na`.
`drop.levels`	Logical, if `TRUE`, unused factor levels will be dropped (i.e. `droplevels` will be applied before returning the result).

Value

A factor with the associated value labels as factor levels. If `x` is a data frame, the complete data frame `x` will be returned, where variables specified in `...` are coerced to factors; if `...` is not specified, applies to all variables in the data frame.

Note

Value label attributes will be removed when converting variables to factors.

This function is kept for backwards-compatibility. It is preferred to use as_label.

Examples

```
library(sjlabelled)
data(efc)
print(get_labels(efc)['c161sex'])
head(efc$c161sex)
head(to_label(efc$c161sex))

print(get_labels(efc)['e42dep'])
table(efc$e42dep)
table(to_label(efc$e42dep))

head(efc$e42dep)
head(to_label(efc$e42dep))

# structure of numeric values won't be changed
# by this function, it only applies to labelled vectors
# (typically categorical or factor variables)
str(efc$e17age)
str(to_label(efc$e17age))


# factor with non-numeric levels
to_label(factor(c("a", "b", "c")))

# factor with non-numeric levels, prefixed
x <- factor(c("a", "b", "c"))
x <- set_labels(x, labels = c("ape", "bear", "cat"))
to_label(x, prefix = TRUE)


# create vector
x <- c(1, 2, 3, 2, 4, NA)
# add less labels than values
x <- set_labels(x,
                labels = c("yes", "maybe", "no"),
                force.labels = FALSE,
                force.values = FALSE)
# convert to label w/o non-labelled values
to_label(x)
# convert to label, including non-labelled values
to_label(x, add.non.labelled = TRUE)


# create labelled integer, with missing flag
library(haven)
x <- labelled(c(1:3, tagged_na("a", "c", "z"), 4:1, 2:3),
```

```
              c("Agreement" = 1, "Disagreement" = 4, "First" = tagged_na("c"),
                "Refused" = tagged_na("a"), "Not home" = tagged_na("z")))
# to labelled factor, with missing labels
to_label(x, drop.na = FALSE)
# to labelled factor, missings removed
to_label(x, drop.na = TRUE)
# keep missings, and use non-labelled values as well
to_label(x, add.non.labelled = TRUE, drop.na = FALSE)


# convert labelled character to factor
dummy <- c("M", "F", "F", "X")
dummy <- set_labels(
  dummy,
  labels = c(`M` = "Male", `F` = "Female", `X` = "Refused")
)
get_labels(dummy,, "p")
to_label(dummy)

# drop unused factor levels, but preserve variable label
x <- factor(c("a", "b", "c"), levels = c("a", "b", "c", "d"))
x <- set_labels(x, labels = c("ape", "bear", "cat"))
set_label(x) <- "A factor!"
x
to_label(x, drop.levels = TRUE)

# change variable label
to_label(x, var.label = "New variable label!", drop.levels = TRUE)


# easily coerce specific variables in a data frame to factor
# and keep other variables, with their class preserved
to_label(efc, e42dep, e16sex, c172code)
```

---

to_long *Convert wide data to long format*

Description

This function converts wide data into long format. It allows to transform multiple key-value pairs to be transformed from wide to long format in one single step.

Usage

```
to_long(data, keys, values, ..., labels = NULL, recode.key = FALSE)
```

Arguments

data	A data.frame that should be tansformed from wide to long format.
keys	Character vector with name(s) of key column(s) to create in output. Either one key value per column group that should be gathered, or a single string. In the latter case, this name will be used as key column, and only one key column is created. See 'Examples'.
values	Character vector with names of value columns (variable names) to create in output. Must be of same length as number of column groups that should be gathered. See 'Examples'.
...	Specification of columns that should be gathered. Must be one character vector with variable names per column group, or a numeric vector with column indices indicating those columns that should be gathered. See 'Examples'.
labels	Character vector of same length as values with variable labels for the new variables created from gathered columns. See 'Examples' and 'Details'.
recode.key	Logical, if TRUE, the values of the key column will be recoded to numeric values, in sequential ascending order.

Details

This function enhances **tidyr**'s gather function that you can gather multiple column groups at once. Value and variable labels for non-gathered variables are preserved. However, gathered variables may have different variable label attributes. In this case, gather will drop these attributes. Hence, the new created variables from gathered columns don't have any variable label attributes. In such cases, use labels argument to set variable label attributes.

Examples

```
# create sample
mydat <- data.frame(age = c(20, 30, 40),
                    sex = c("Female", "Male", "Male"),
                    score_t1 = c(30, 35, 32),
                    score_t2 = c(33, 34, 37),
                    score_t3 = c(36, 35, 38),
                    speed_t1 = c(2, 3, 1),
                    speed_t2 = c(3, 4, 5),
                    speed_t3 = c(1, 8, 6))

# check tidyr. score is gathered, however, speed is not
tidyr::gather(mydat, "time", "score", score_t1, score_t2, score_t3)

# gather multiple columns. both time and speed are gathered.
to_long(
  data = mydat,
  keys = "time",
  values = c("score", "speed"),
  c("score_t1", "score_t2", "score_t3"),
  c("speed_t1", "speed_t2", "speed_t3")
)
```

```
# gather multiple columns, use numeric key-value
to_long(
  data = mydat,
  keys = "time",
  values = c("score", "speed"),
  c("score_t1", "score_t2", "score_t3"),
  c("speed_t1", "speed_t2", "speed_t3"),
  recode.key = TRUE
)

# gather multiple columns by colum names and colum indices
to_long(
  data = mydat,
  keys = "time",
  values = c("score", "speed"),
  c("score_t1", "score_t2", "score_t3"),
  6:8,
  recode.key = TRUE
)

# gather multiple columns, use separate key-columns
# for each value-vector
to_long(
  data = mydat,
  keys = c("time_score", "time_speed"),
  values = c("score", "speed"),
  c("score_t1", "score_t2", "score_t3"),
  c("speed_t1", "speed_t2", "speed_t3")
)

# gather multiple columns, label columns
mydat <- to_long(
  data = mydat,
  keys = "time",
  values = c("score", "speed"),
  c("score_t1", "score_t2", "score_t3"),
  c("speed_t1", "speed_t2", "speed_t3"),
  labels = c("Test Score", "Time needed to finish")
)

library(sjlabelled)
str(mydat$score)
get_label(mydat$speed)
```

---

to_value *Convert factors to numeric variables*

Description

This function converts (replaces) factor levels with the related factor level index number, thus the factor is converted to a numeric variable.

Usage

```
to_value(x, ..., start.at = NULL, keep.labels = TRUE)
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
start.at	Starting index, i.e. the lowest numeric value of the variable's value range. By default, this argument is NULL, hence the lowest value of the returned numeric variable corresponds to the lowest factor level (if factor levels are numeric) or to 1 (if factor levels are not numeric).
keep.labels	Logical, if TRUE, former factor levels will be added as value labels. For numeric factor levels, values labels will be used, if present. See 'Examples' and set_labels for more details.

Value

A numeric variable with values ranging either from start.at to start.at + length of factor levels, or to the corresponding factor levels (if these were numeric). If x is a data frame, the complete data frame x will be returned, where variables specified in ... are coerced to numeric; if ... is not specified, applies to all variables in the data frame.

Note

This function is kept for backwards-compatibility. It is preferred to use as_numeric.

Examples

```
data(efc)
test <- to_label(efc$e42dep)
table(test)

table(to_value(test))
hist(to_value(test, start.at = 0))

# set lowest value of new variable to "5".
table(to_value(test, start.at = 5))

# numeric factor keeps values
dummy <- factor(c("3", "4", "6"))
table(to_value(dummy))
```

```
# do not drop unused factor levels
dummy <- ordered(c(rep("No", 5), rep("Maybe", 3)),
                 levels = c("Yes", "No", "Maybe"))
to_value(dummy)

# non-numeric factor is converted to numeric
# starting at 1
dummy <- factor(c("D", "F", "H"))
table(to_value(dummy))

library(sjlabelled)
# for numeric factor levels, value labels will be used, if present
dummy1 <- factor(c("3", "4", "6"))
dummy1 <- set_labels(dummy1, labels = c("first", "2nd", "3rd"))
dummy1
to_value(dummy1)

# for non-numeric factor levels, these will be used.
# value labels will be ignored
dummy2 <- factor(c("D", "F", "H"))
dummy2 <- set_labels(dummy2, labels = c("first", "2nd", "3rd"))
dummy2
to_value(dummy2)


# easily coerce specific variables in a data frame to numeric
# and keep other variables, with their class preserved
data(efc)
efc$e42dep <- as.factor(efc$e42dep)
efc$e16sex <- as.factor(efc$e16sex)
efc$e17age <- as.factor(efc$e17age)

# convert back "sex" and "age" into numeric
to_value(efc, e16sex, e17age)
```

`trim` *Trim leading and trailing whitespaces from strings*

Description

Trims leading and trailing whitespaces from strings or character vectors.

Usage

```
trim(x)
```

Arguments

x	Character vector or string, or a list or data frame with such vectors. Function is vectorized, i.e. vector may have a length greater than 1. See 'Examples'.

Value

Trimmed x, i.e. with leading and trailing spaces removed.

Examples

```
trim("white space at end ")
trim(" white space at start and end ")
trim(c(" string1 ", "   string2", "string 3   "))

tmp <- data.frame(a = c(" string1 ", "   string2", "string 3   "),
                  b = c(" strong one  ", "    string two", "  third string "),
                  c = c(" str1 ", "   str2", "str3   "))
tmp
trim(tmp)
```

---

var_rename *Rename variables*

Description

This function renames variables in a data frame, i.e. it renames the columns of the data frame.

Usage

```
var_rename(x, ...)
```

Arguments

x	A data frame.
...	Pairs of named vectors, where the name equals the column name that should be renamed, and the value is the new column name.

Value

x, with new column names for those variables specified in ....

Examples

```
# Set variable labels for data frame
dummy <- data.frame(a = sample(1:4, 10, replace = TRUE),
                    b = sample(1:4, 10, replace = TRUE),
                    c = sample(1:4, 10, replace = TRUE))

var_rename(dummy, a = "first.col", c = "3rd.col")
```

---

var_type *Determine variable type*

Description

This function returns the type of a variable as character. It is similar to type_sum, however, the return value is not truncated, and var_type() works on data frames and within pipe-chains.

Usage

```
var_type(x, ..., abbr = FALSE)
```

Arguments

x	A vector or data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.
abbr	Logical, if TRUE, returns a shortened, abbreviated value for the variable type (as returned by type_sum). If FALSE (default), a longer "description" is returned.

Value

The variable type of x, as character.

Examples

```
data(efc)

var_type(1)
var_type(1L)
var_type("a")

var_type(efc$e42dep)
var_type(to_factor(efc$e42dep))

library(dplyr)
var_type(efc, contains("cop"))
```

word_wrap *Insert line breaks in long labels*

Description

Insert line breaks in long character strings. Useful if you want to wordwrap labels / titles for plots or tables.

Usage

```
word_wrap(labels, wrap, linesep = NULL)
```

Arguments

labels	Label(s) as character string, where a line break should be inserted. Several strings may be passed as vector (see 'Examples').
wrap	Maximum amount of chars per line (i.e. line length). If `wrap = Inf`, no word wrap will be performed (i.e. `labels` will be returned as is).
linesep	By default, this argument is `NULL` and a regular new line string (`"\n"`) is used. For HTML-purposes, for instance, `linesep` could be `" "`.

Value

New label(s) with line breaks inserted at every `wrap`'s position.

Examples

```
word_wrap(c("A very long string", "And another even longer string!"), 10)

message(word_wrap("Much too long string for just one line!", 15))
```

zap_inf *Convert infiite or NaN values into regular NA*

Description

Replaces all infinite (`Inf` and `-Inf`) or `NaN` values with regular `NA`.

Usage

```
zap_inf(x, ...)
```

Arguments

x	A vector or a data frame.
...	Optional, unquoted names of variables that should be selected for further processing. Required, if x is a data frame (and no vector) and only selected variables from x should be processed. You may also use functions like : or tidyselect's select_helpers. See 'Examples' or package-vignette.

Value

x, where all Inf, -Inf and NaN are converted to NA.

Examples

```
x <- c(1, 2, NA, 3, NaN, 4, NA, 5, Inf, -Inf, 6, 7)
zap_inf(x)

data(efc)
# produce some NA and NaN values
efc$e42dep[1] <- NaN
efc$e42dep[2] <- NA
efc$c12hour[1] <- NaN
efc$c12hour[2] <- NA
efc$e17age[2] <- NaN
efc$e17age[1] <- NA

# only zap NaN for c12hour
zap_inf(efc$c12hour)

# only zap NaN for c12hour and e17age, not for e42dep,
# but return complete data framee
zap_inf(efc, c12hour, e17age)

# zap NaN for complete data frame
zap_inf(efc)
```

%nin% *Value matching*

Description

%nin% is the complement to %in%. It looks which values in x do *not* match (hence, are *not in*) values in y.

Usage

```
x %nin% y
```

Arguments

x	Vector with values to be matched.
y	Vector with values to be matched against.

Details

See 'Details' in match.

Value

A logical vector, indicating if a match was *not* located for each element of x, thus the values are TRUE or FALSE and never NA.

Examples

```
c("a", "B", "c") %in% letters
c("a", "B", "c") %nin% letters

c(1, 2, 3, 4) %in% c(3, 4, 5, 6)
c(1, 2, 3, 4) %nin% c(3, 4, 5, 6)
```

Index