Package 'expss'

January 24, 2018

Type Package

Title Tables with Labels and Some Useful Functions from Spreadsheets and 'SPSS' Statistics

Version 0.8.6

Date 2018-01-24

Maintainer Gregory Demin <gdemin@gmail.com>

URL https://gdemin.github.io/expss/

BugReports https://github.com/gdemin/expss/issues

Depends R (>= 3.3.0),

Imports foreign, utils, stats, magrittr (>= 1.5), htmlTable (>= 1.11.0), matrixStats (>= 0.51.0), data.table (>= 1.10),

Suggests DT, htmltools, knitr, repr, ggplot2, testthat

Description Package provides tabulation functions with support for 'SPSS'-style labels, multiple / nested banners, weights, multiple-response variables and significance testing. There are facilities for nice output of tables in 'knitr', R notebooks, 'Shiny' and 'Jupyter' notebooks. Proper methods for labelled variables add value labels support to base R functions and to some functions from other packages. Additionally, the package offers useful functions for data processing in marketing research / social surveys - popular data transformation functions from 'SPSS' Statistics ('RECODE', 'COUNT', 'COMPUTE', 'DO IF', etc.) and 'Excel' ('COUNTIF', 'VLOOKUP', etc.). Package is intended to help people to move data processing from 'Excel'/'SPSS' to R.

VignetteBuilder knitr

LazyData yes

License GPL (>= 2)

RoxygenNote 6.0.1

NeedsCompilation no

Author Gregory Demin [aut, cre]

Repository CRAN

Date/Publication 2018-01-24 16:28:03 UTC

R topics documented:

`add_rows` *Add rows to data.frame/matrix/table*

Description

`add_rows` is similar to rbind but it handles non-matching column names. `%add_rows%` is an infix version of `add_rows`. There is also special method for the results of cro/cro_fun/tables/fre. `.add_rows` is version for adding rows to default dataset. See default_dataset.

Usage

```
add_rows(..., nomatch_columns = c("add", "drop", "stop"))

## S3 method for class 'data.frame'
add_rows(..., nomatch_columns = c("add", "drop", "stop"))

x %add_rows% y

.add_rows(..., nomatch_columns = c("add", "drop", "stop"))
```

Arguments

`...`	data.frame/matrix/table for binding
`nomatch_columns`	action if there are non-matching columns between data.frames. Possible values are `"add"`, `"drop"`, `"stop"`. `"add"` will combine all columns, `"drop"` will leave only common columns, `"stop"` will raise an error.
`x`	data.frame/matrix/table for binding
`y`	data.frame/matrix/table for binding

Value

See rbind, cro, cro_fun, fre, tables

Examples

```
a = data.frame(x = 1:5, y = 6:10)
b = data.frame(y = 6:10, z = 11:15)

add_rows(a, b) # x, y, z
a %add_rows% b # the same result

add_rows(a, b, nomatch_columns = "drop")  # y

# simple tables
data(mtcars)
# apply labels
mtcars = apply_labels(mtcars,
                mpg = "Miles/(US) gallon",
                cyl = "Number of cylinders",
                disp = "Displacement (cu.in.)",
                hp = "Gross horsepower",
                drat = "Rear axle ratio",
                wt = "Weight (lb/1000)",
                qsec = "1/4 mile time",
                vs = "V/S",
                vs = c("V-engine" = 0, "Straight engine" = 1),
                am = "Transmission (0 = automatic, 1 = manual)",
                am = c(automatic = 0, manual = 1),
                gear = "Number of forward gears",
                carb = "Number of carburetors"
)

tbl_mean = calculate(mtcars, cro_mean(cyl, am))
tbl_percent = calculate(mtcars, cro_cpct(cyl, am))

tbl_mean %add_rows% tbl_percent
```

---

apply_labels *Set variable labels/value labels on variables in the data.frame*

Description

`apply_labels` tries automatically detect what is variable label and what are value labels. `.apply_*` are versions for working with default dataset. See also var_lab and val_lab.

Usage

```
apply_labels(data, ...)

apply_var_labs(data, ...)

apply_val_labs(data, ...)

.apply_labels(...)

.apply_val_labs(...)

.apply_var_labs(...)
```

Arguments

data	data.frame/list
...	named arguments. Name of argument is a variable name in data. Arguments values is variable label/value labels. For apply_labels unnamed characters of length 1 are considered as variable labels and named vectors are considered as value labels.

Value

data with applied labels

Examples

```
data(mtcars)
mtcars = apply_labels(mtcars,
                      vs = "Engine",
                      vs = num_lab("
                             0 V-engine
                             1 Straight engine
                             "),
                      am = "Transmission",
                      am = num_lab("
                             0 Automatic
                             1 Manual
                             ")
)

# 'table' from base R
table(mtcars$vs, mtcars$am)

# more sofisticated crosstable
calculate(mtcars, cro(vs, am))
```

`as.category` *Convert dichotomy data.frame/matrix to data.frame with category encoding*

Description

Convert dichotomy data.frame/matrix to data.frame with category encoding

Usage

```
as.category(x, prefix = NULL, counted_value = 1, compress = FALSE)

is.category(x)
```

Arguments

`x`	Dichotomy data.frame/matrix (usually with 0,1 coding).
`prefix`	If is not NULL then column names will be added in the form prefix+column number.
`counted_value`	Vector. Values that will be considered as indicator of category presence. By default it equals to 1.
`compress`	Logical. Should we drop columns with all NA? FALSE by default. TRUE significantly decreases performance of the function.

Value

data.frame of class `category` with numeric values that correspond to column numbers of counted values. Column names of x or variable labels are added as value labels.

See Also

`as.dichotomy` for reverse conversion, mrset, mdset for usage multiple-response variables with tables.

Examples

```
set.seed(123)

# Let's imagine it's matrix of consumed products
dichotomy_matrix = matrix(sample(0:1,40,replace = TRUE,prob=c(.6,.4)),nrow=10)
colnames(dichotomy_matrix) = c("Milk","Sugar","Tea","Coffee")

as.category(dichotomy_matrix, compress = TRUE) # compressed version
category_encoding = as.category(dichotomy_matrix)

 # should be TRUE
identical(val_lab(category_encoding), c(Milk = 1L, Sugar = 2L, Tea = 3L, Coffee = 4L))
all(as.dichotomy(category_encoding, use_na = FALSE) == dichotomy_matrix)
```

```
# with prefix
as.category(dichotomy_matrix, prefix = "products_")

# data.frame with variable labels
dichotomy_dataframe = as.data.frame(dichotomy_matrix)
colnames(dichotomy_dataframe) = paste0("product_", 1:4)
var_lab(dichotomy_dataframe[[1]]) = "Milk"
var_lab(dichotomy_dataframe[[2]]) = "Sugar"
var_lab(dichotomy_dataframe[[3]]) = "Tea"
var_lab(dichotomy_dataframe[[4]]) = "Coffee"

as.category(dichotomy_dataframe, prefix = "products_")
```

`as.datatable_widget` *Create an HTML table widget for usage with Shiny*

Description

This is method for rendering results of tables/fre/cro in Shiny. `DT` package should be installed for this feature (`install.packages('DT')`). For detailed description of function and its arguments see datatable.

Usage

```
as.datatable_widget(data, ...)

datatable(data, ...)

## S3 method for class 'etable'
as.datatable_widget(data, repeat_row_labels = FALSE,
  show_row_numbers = FALSE, digits = get_expss_digits(), ...)
```

Arguments

`data`	a data object (result of tables/fre/cro).
`...`	further parameters for datatable
`repeat_row_labels`	logical Should we repeat duplicated row labels in the every row? Default is FALSE.
`show_row_numbers`	logical Default is FALSE.
`digits`	integer By default, all numeric columns are rounded to one digit after decimal separator. Also you can set this argument by option 'expss.digits' - for example, `expss_digits(2)`. If it is NA than all numeric columns remain unrounded.

Value

Object of class datatable

See Also

htmlTable for knitting

Examples

```
## Not run:

data(mtcars)
mtcars = apply_labels(mtcars,
                      mpg = "Miles/(US) gallon",
                      cyl = "Number of cylinders",
                      disp = "Displacement (cu.in.)",
                      hp = "Gross horsepower",
                      drat = "Rear axle ratio",
                      wt = "Weight (1000 lbs)",
                      qsec = "1/4 mile time",
                      vs = "Engine",
                      vs = c("V-engine" = 0,
                             "Straight engine" = 1),
                      am = "Transmission",
                      am = c("Automatic" = 0,
                             "Manual"=1),
                      gear = "Number of forward gears",
                      carb = "Number of carburetors"
)

mtcars_table = mtcars %>%
                  tab_cols(total(), am %nest% vs) %>%
                  tab_cells(mpg, hp) %>%
                  tab_stat_mean() %>%
                  tab_cells(cyl) %>%
                  tab_stat_cpct() %>%
                  tab_pivot()

library(shiny)
shinyApp(
    ui = fluidPage(fluidRow(column(12, DT::dataTableOutput('tbl')))),
    server = function(input, output) {
        output$tbl = DT::renderDataTable(
            as.datatable_widget(mtcars_table)
        )
    }
)

## End(Not run)
```

`as.dichotomy` *Convert variable (possibly multiple choice question) to data.frame/matrix of dummy variables.*

Description

This function converts variable/multiple response variable (vector/matrix/data.frame) with category encoding into data.frame/matrix with dichotomy encoding (0/1) suited for most statistical analysis, e. g. clustering, factor analysis, linear regression and so on.

- `as.dichotomy` returns data.frame of class 'dichotomy' with 0, 1 and possibly NA.
- `dummy` returns matrix of class 'dichotomy' with 0, 1 and possibly NA.
- `dummy1` drops last column in dichotomy matrix. It is useful in many cases because any column of such matrix usually is linear combinations of other columns.

Usage

```
as.dichotomy(x, prefix = "v", keep_unused = FALSE, use_na = TRUE,
  keep_values = NULL, keep_labels = NULL, drop_values = NULL,
  drop_labels = NULL, presence = 1, absence = 0)

dummy(x, keep_unused = FALSE, use_na = TRUE, keep_values = NULL,
  keep_labels = NULL, drop_values = NULL, drop_labels = NULL,
  presence = 1, absence = 0)

dummy1(x, keep_unused = FALSE, use_na = TRUE, keep_values = NULL,
  keep_labels = NULL, drop_values = NULL, drop_labels = NULL,
  presence = 1, absence = 0)

is.dichotomy(x)
```

Arguments

`x`	vector/factor/matrix/data.frame.
`prefix`	character. By default "v".
`keep_unused`	Logical. Should we create columns for unused value labels/factor levels? FALSE by default.
`use_na`	Logical. Should we use NA for rows with all NA or use 0's instead. TRUE by default.
`keep_values`	Numeric/character. Values that should be kept. By default all values will be kept.
`keep_labels`	Numeric/character. Labels/levels that should be kept. By default all labels/levels will be kept.
`drop_values`	Numeric/character. Values that should be dropped. By default all values will be kept. Ignored if keep_values/keep_labels are provided.

drop_labels	Numeric/character. Labels/levels that should be dropped. By default all labels/levels will be kept. Ignored if keep_values/keep_labels are provided.
presence	numeric value which will code presence of the level. By default it is 1. Note that all tables functions need that presence and absence will be 1 and 0.
absence	numeric value which will code absence of the level. By default it is 0. Note that all tables functions need that presence and absence will be 1 and 0.

Value

as.dichotomy returns data.frame of class dichotomy with 0,1. Columns of this data.frame have variable labels according to value labels of original data. If label doesn't exist for particular value then this value will be used as variable label. dummy returns matrix of class dichotomy. Column names of this matrix are value labels of original data.

See Also

as.category for reverse conversion, mrset, mdset for usage multiple-response variables with tables.

Examples

```
# toy example
# brands - multiple response question
# Which brands do you use during last three months?
set.seed(123)
brands = as.dtfrm(t(replicate(20,sample(c(1:5,NA),4,replace = FALSE))))
# score - evaluation of tested product
score = sample(-1:1,20,replace = TRUE)
var_lab(brands) = "Used brands"
val_lab(brands) = autonum("
                          Brand A
                          Brand B
                          Brand C
                          Brand D
                          Brand E
                          ")

var_lab(score) = "Evaluation of tested brand"
val_lab(score) = make_labels("
                         -1 Dislike it
                          0 So-so
                          1 Like it
                         ")

cro_cpct(as.dichotomy(brands), score)
# the same as
cro_cpct(mrset(brands), score)

# customer segmentation by used brands
kmeans(dummy(brands), 3)
```

```
# model of influence of used brands on evaluation of tested product
summary(lm(score ~ dummy(brands)))

# prefixed data.frame
as.dichotomy(brands, prefix = "brand_")
```

`as.etable` *Convert data.frame/matrix to object of class 'etable'*

Description

If `x` is `data.frame` then `as.etable` just adds `etable` to `class` attribute of `x`. If `x` is matrix then it will be converted to data.frame.

Usage

```
as.etable(x, rownames_as_row_labels = NULL)

is.etable(x)
```

Arguments

`x`	data.frame/matrix
`rownames_as_row_labels`	logical. If it is TRUE than rownames of `x` will be added to result as first column with name `row_labels`. By default row names will be added if they are not NULL and are not sequential numerics.

Value

object of class `etable`

Examples

```
data(mtcars)
etable_mtcars = as.etable(mtcars)
is.etable(etable_mtcars) #TRUE

etable_mtcars #another 'print' method is used

cor(mtcars) %>% as.etable()
```

as.labelled *Recode vector into numeric vector with value labels*

Description

Recode vector into numeric vector with value labels

Usage

```
as.labelled(x, label = NULL)

is.labelled(x)
```

Arguments

x	numeric vector/character vector/factor
label	optional variable label

Value

numeric vector with labels

Examples

```
character_vector = c("one", "two",  "two", "three")
as.labelled(character_vector, label = "Numbers")
```

by_groups *Aggregate dataset by grouping variable(s).*

Description

Splits the data by groups, computes summary statistics for each, and returns data.frame/data.table.

Usage

```
by_groups(data, ...)
```

Arguments

data	data for aggregation
...	aggregation parameters. Character/numeric or criteria/logical functions (see criteria) for grouping variables. Names of variables at the top-level can be unquoted (non-standard evaluation). For standard evaluation of parameters, you can surround them by round brackets. You need additionally specify formulas with aggregation expressions, such as `mean_x ~ mean(x)`. Instead of the formulas it can be single function as last argument - it will be applied to all non-grouping columns. See examples.

Value

aggregated data.frame/data.table

Examples

```
# compute mean of the every column for every value of the Species
data(iris)
by_groups(iris, Species, mean)

# compute mean of the every numeric column
iris %>% except(Species) %>% by_groups(mean)

# compute different functions for different columns
# automatic naming
data(mtcars)
by_groups(mtcars, cyl, am, ~ mean(hp), ~ median(mpg))

# with custom names
by_groups(mtcars, cyl, am, mean_hp ~ mean(hp), median_mpg ~ median(mpg))

# variable substitution
group1 = "cyl"
statistic1 = ~ mean(hp)
by_groups(mtcars, (group1), (statistic1))

group2 = "am"
# formulas can be easily constructed from text strings
statistic2 = as.formula("~ median(mpg)")
by_groups(mtcars, (group2), (statistic2))

by_groups(mtcars, (group1), (group2), (statistic1), (statistic2))
```

`compare_proportions` *Calculate significance (p-values) of differences between proportions/means*

Description

compare_proportions calculates p-values (via z-test) for comparison between each proportion in the prop1 and prop2. Results are calculated with the same formula as in prop.test without continuity correction. compare_means calculates p-values (via t-test) for comparison between each mean in the mean1 and mean2. Results are calculated on the aggregated statistics (means, std. devs, N) with the same formula as in t.test. These functions mainly intended for usage inside significance_cpct and significance_means.

Usage

```
compare_proportions(prop1, prop2, base1, base2, common_base = 0)

compare_means(mean1, mean2, sd1, sd2, base1, base2, common_base = 0,
  var_equal = FALSE)
```

Arguments

prop1	a numeric vector of proportions in the group 1. Values should be between 0 and 1
prop2	a numeric vector of proportions in the group 2. Values should be between 0 and 1
base1	a numeric vector for compare_means and single number for compare_proportions. Number of valid cases for each mean in the first group for compare_means and number of cases for compare_proportions.
base2	a numeric vector for compare_means and single number for compare_proportions. Number of valid cases for each mean in the second group for compare_means and number of cases for compare_proportions.
common_base	numeric. Number of cases that belong to both values in the first and the second argument. It can occur in the case of overlapping samples. Calculations are made according to algorithm in IBM SPSS Statistics Algorithms v20, p. 263. Note that with these adjustments t-tests between means are made with equal variance assumed (as with var_equal = TRUE).
mean1	a numeric vector of the means in the first group.
mean2	a numeric vector of the means in the second group.
sd1	a numeric vector of the standard deviations in the first group. Values should be non-negative.
sd2	a numeric vector of the standard deviations in the second group. Values should be non-negative.
var_equal	a logical variable indicating whether to treat the variances in the groups as being equal. For details see t.test.

Value

numeric vector with p-values

See Also

significance_cpct, significance_means, prop.test, t.test

Examples

```
# proportions
data(mtcars)
counts = table(mtcars$am, mtcars$vs)
props = prop.table(counts)
compare_proportions(props[,1], props[,2],
                    colSums(counts)[1], colSums(counts)[1])

# means
t.test(mpg ~ am, data = mtcars)$p.value
# the same result
calculate(mtcars,
          compare_means(
              mean(mpg[am==0]), mean(mpg[am==1]),
              sd(mpg[am==0]),  sd(mpg[am==1]),
              length(mpg[am==0]), length(mpg[am==1])
          ))
```

---

count_if *Count/sum/average/other functions on values that meet a criterion*

Description

These functions calculate count/sum/average/etc on values that meet a criterion that you specify. `apply_if_*` apply custom functions. There are different flavors of these functions: `*_if` work on entire dataset/matrix/vector, `*_row_if` works on each row and `*_col_if` works on each column.

Usage

```
count_if(criterion, ...)

count_row_if(criterion, ...)

count_col_if(criterion, ...)

x %row_in% criterion

x %has% criterion

x %col_in% criterion

sum_if(criterion, ..., data = NULL)

sum_row_if(criterion, ..., data = NULL)
```

```
sum_col_if(criterion, ..., data = NULL)

mean_if(criterion, ..., data = NULL)

mean_row_if(criterion, ..., data = NULL)

mean_col_if(criterion, ..., data = NULL)

sd_if(criterion, ..., data = NULL)

sd_row_if(criterion, ..., data = NULL)

sd_col_if(criterion, ..., data = NULL)

median_if(criterion, ..., data = NULL)

median_row_if(criterion, ..., data = NULL)

median_col_if(criterion, ..., data = NULL)

max_if(criterion, ..., data = NULL)

max_row_if(criterion, ..., data = NULL)

max_col_if(criterion, ..., data = NULL)

min_if(criterion, ..., data = NULL)

min_row_if(criterion, ..., data = NULL)

min_col_if(criterion, ..., data = NULL)

apply_row_if(fun, criterion, ..., data = NULL)

apply_col_if(fun, criterion, ..., data = NULL)
```

Arguments

criterion	Vector with counted values, logical vector/matrix or function. See details and examples.
...	Data on which criterion will be applied. Vector, matrix, data.frame, list. Shorter arguments will be recycled.
x	Data on which criterion will be applied. Vector, matrix, data.frame, list. Shorter arguments will be recycled.
data	Data on which function will be applied. Doesn't applicable to `count_*_if` functions. If omitted then function will be applied on the ... argument.
fun	Custom function that will be applied based on criterion.

Details

Possible type for criterion argument:

- vector/single value All values in `...` which equal to elements of vector in criteria will be used as function `fun` argument.
- function Values for which function gives TRUE will be used as function `fun` argument. There are some special functions for convenience (e. g. `gt(5)` is equivalent ">5" in spreadsheet) - see criteria.
- logical vector/matrix/data.frame Values for which element of criterion equals to TRUE will be used as function `fun` argument. Logical vector will be recycled across all columns of `...data`. If criteria is logical matrix/data.frame then column from this matrix/data.frame will be used for corresponding column/element of `...data`. Note that this kind of criterion doesn't use `...` so `...` can be used instead of `data` argument.

`count*` and `%in*%` never returns NA's. Other functions remove NA's before calculations (as `na.rm = TRUE` in base R functions).

Function criterion should return logical vector of same size and shape as its argument. This function will be applied to each column of supplied data and TRUE results will be used. There is asymmetrical behavior in `*_row_if` and `*_col_if` for function criterion: in both cases function criterion will be applied columnwise.

Value

`*_if` return single value (vector of length 1). `*_row_if` returns vector for each row of supplied arguments. `*_col_if` returns vector for each column of supplied arguments. `%row_in%/%col_in%` return logical vector - indicator of presence of criterion in each row/column. `%has%` is an alias for `%row_in%`.

Examples

```
set.seed(123)
dfs = as.data.frame(
       matrix(sample(c(1:10,NA), 30, replace = TRUE), 10)
)

result  = modify(dfs, {
             # count 8
             exact = count_row_if(8, V1, V2, V3)
             # count values greater than 8
             greater = count_row_if(gt(8), V1, V2, V3)
             # count integer values between 5 and 8, e. g. 5, 6, 7, 8
             integer_range = count_row_if(5:8, V1, V2, V3)
             # count values between 5 and 8
             range = count_row_if(5 %thru% 8, V1, V2, V3)
             # count NA
             na = count_row_if(is.na, V1, V2, V3)
             # count not-NA
             not_na = count_row_if(not_na, V1, V2, V3)
             # are there any 5 in each row?
             has_five = cbind(V1, V2, V3) %row_in% 5
```

```
        })
result

mean_row_if(6, dfs$V1, data = dfs)
median_row_if(gt(2), dfs$V1, dfs$V2, dfs$V3)
sd_row_if(5 %thru% 8, dfs$V1, dfs$V2, dfs$V3)

if_na(dfs) = 5 # replace NA

# custom apply
apply_col_if(prod, gt(2), dfs$V1, data = dfs) # product of all elements by columns
apply_row_if(prod, gt(2), dfs$V1, data = dfs) # product of all elements by rows

# Examples borrowed from Microsoft Excel help for COUNTIF
df1 = data.frame(
    a = c("apples",   "oranges",     "peaches",     "apples"),
    b = c(32, 54, 75, 86)
)

count_if("apples", df1$a) # 2

count_if("apples", df1) # 2

with(df1, count_if("apples", a, b)) # 2

count_if(gt(55), df1$b) # greater than 55 = 2

count_if(ne(75), df1$b) # not equal 75 = 3

count_if(ge(32), df1$b) # greater than or equal 32 = 4

count_if(gt(32) & lt(86), df1$b) # 2

# count only integer values between 33 and 85
count_if(33:85, df1$b) # 2

# values with letters
count_if(regex("^[A-z]+$"), df1) # 4

# values that started on 'a'
count_if(regex("^a"), df1) # 2

# count_row_if
count_row_if(regex("^a"), df1) # c(1,0,0,1)

df1 %row_in% 'apples'  # c(TRUE,FALSE,FALSE,TRUE)

# Some of Microsoft Excel examples for SUMIF/AVERAGEIF/etc
dfs = read.csv(
    text = "
    property_value,commission,data
    100000,7000,250000
    200000,14000,
```

```
    300000,21000,
    400000,28000,"
)

# Sum of commision for property value greater than 160000
with(dfs, sum_if(gt(160000), property_value, data = commission)) # 63000

# Sum of property value greater than 160000
with(dfs, sum_if(gt(160000), property_value)) # 900000

# Sum of commision for property value equals to 300000
with(dfs, sum_if(300000, property_value, data = commission)) # 21000

# Sum of commision for property value greater than first value of data
with(dfs, sum_if(gt(data[1]), property_value, data = commission)) # 49000


dfs = data.frame(
    category = c("Vegetables", "Vegetables", "Fruits", "", "Vegetables", "Fruits"),
    food = c("Tomatoes", "Celery", "Oranges", "Butter", "Carrots", "Apples"),
    sales = c(2300, 5500, 800, 400, 4200, 1200),
    stringsAsFactors = FALSE
)

# Sum of sales for Fruits
with(dfs, sum_if("Fruits", category, data = sales)) # 2000

# Sum of sales for Vegetables
with(dfs, sum_if("Vegetables", category, data = sales)) # 12000

# Sum of sales for food which is ending on 'es'
with(dfs, sum_if(perl("es$"), food, data = sales)) # 4300

# Sum of sales for empty category
with(dfs, sum_if("", category, data = sales))  # 400


dfs = read.csv(
    text = "
    property_value,commission,data
    100000,7000,250000
    200000,14000,
    300000,21000,
    400000,28000,"
)

# Commision average for comission less than 23000
with(dfs, mean_if(lt(23000), commission)) # 14000


# Property value average for property value less than 95000
with(dfs, mean_if(lt(95000), property_value)) #  NaN
```

```
# Commision average for property value greater than 250000
with(dfs, mean_if(gt(250000), property_value, data = commission)) # 24500


dfs = data.frame(
    region = c("East", "West", "North", "South (New Office)",  "MidWest"),
    profits = c(45678, 23789, -4789, 0, 9678),
    stringsAsFactors = FALSE
)


# Mean profits for 'west' regions
with(dfs, mean_if(fixed("West"), region, data = profits)) # 16733.5


# Mean profits for regions wich doesn't contain New Office
with(dfs, mean_if(!fixed("(New Office)"), region, data = profits))  # 18589


dfs = read.csv(
    text = '
    grade,weight
    89,1
    93,2
    96,2
    85,3
    91,1
    88,1'
    ,stringsAsFactors = FALSE
)

# Minimum grade for weight equals to 1
with(dfs, min_if(1, weight, data = grade)) # 88


# Maximum grade for weight equals to 1
with(dfs, max_if(1, weight, data = grade)) #91


# Example with offset
dfs = read.csv(
    text = '
    weight,grade
    10,b
    11,a
    100,a
    111,b
    1,a
    1,a'
    ,stringsAsFactors = FALSE
)

with(dfs, min_if("a", grade[2:5], data = weight[1:4])) # 10
```

`criteria` *Criteria functions*

Description

These functions returns criteria functions which could be used in different situation - see keep, except, recode, na_if, %i%, %d%, count_if, match_row etc. For example, `gt(5)` returns function which tests whether its argument greater than five. `fixed("apple")` return function which tests whether its argument contains "apple". Logical operations (|, &, !, xor) are defined for these functions. List of functions:

- `gt` greater than
- `ge`/`gte` greater than or equal
- `eq` equal
- `ne`/`neq` not equal
- `lt` less than
- `le`/`lte` less than or equal
- `thru` checks whether value is inside interval. `thru(0,1)` is equivalent of `x>=0 & x<=1` or `ge(0) & le(1)`
- `%thru%` infix version of `thru`, e. g. `0 %thru% 1`
- `regex` use POSIX 1003.2 extended regular expressions. For details see grepl
- `perl` perl-compatible regular expressions. For details see grepl
- `fixed` pattern is a string to be matched as is. For details see grepl
- `to` returns function which gives TRUE for all elements of vector before the first occurrence of `x` and for `x`.
- `from` returns function which gives TRUE for all elements of vector after the first occurrence of `x` and for `x`. `from` and `to` are intended for usage with keep and except.
- `not_na` return TRUE for all non-NA elements of vector.
- `other` return TRUE for all elements of vector. It is intended for usage with `if_val`, `keep`, `except`
- `items` return TRUE for elements of vector with given sequential number. It is intended for usage with `keep`, `except`

Usage

```
eq(x)

ne(x)

neq(x)

lt(x)

gt(x)

le(x)

lte(x)

ge(x)

gte(x)

perl(pattern, ignore.case = FALSE, useBytes = FALSE)

regex(pattern, ignore.case = FALSE, useBytes = FALSE)

fixed(pattern, ignore.case = FALSE, useBytes = FALSE)

thru(lower, upper)

lower %thru% upper

from(x)

to(x)

items(...)

not_na(x)

other(x)

as.criterion(crit)
```

Arguments

x	vector
pattern	character string containing a regular expression (or character string for `fixed`) to be matched in the given character vector. Coerced by as.character to a character string if possible.
ignore.case	logical see grepl

useBytes	logical see grepl
lower	vector/single value - lower bound of interval
upper	vector/single value - upper bound of interval
...	numeric indexes of desired items
crit	vector of values/function which returns logical or vector. It will be converted to function of class criterion.

Value

function of class 'criterion' which tests its argument against condition and return logical value

See Also

recode, keep, except, count_if, match_row, na_if, %i%, %d%

Examples

```
# operations on vector
1:6 %d% gt(4) # 1:4

1:6 %d% (1 | gt(4)) # 2:4

letters %i% (fixed("a") | fixed("z")) # a, z

letters %i% from("w")  # w, x, y, z

letters %i% to("c")  # a, b, c

letters %i% (from("b") & to("e"))  # b, d, e

c(1, 2, NA, 3) %i% other # c(1, 2, 3)

# examples with count_if
df1 = data.frame(
    a=c("apples", "oranges", "peaches", "apples"),
    b = c(32, 54, 75, 86)
)

count_if(gt(55), df1$b) # greater than 55 = 2

count_if(ne(75), df1$b) # not equal 75 = 3

count_if(ge(32), df1$b) # greater than or equal 32 = 4

count_if(gt(32) & lt(86), df1$b) # greater than 32 and less than 86 = 2

# via different kinds of 'thru'
count_if(thru(35, 80), df1$b) # greater than or equals to 35 and less than or equals to 80 = 2
# infix version
count_if(35 %thru% 80, df1$b) # greater than or equals to 35 and less than or equals to 80 = 2
```

```
# values that started on 'a'
count_if(regex("^a"), df1) # 2

# count_row_if
count_row_if(regex("^a"), df1) # c(1,0,0,1)

# examples with 'keep' and 'except'

data(iris)
iris %>% keep(to("Petal.Width")) # column 'Species' will be removed

# 'Sepal.Length', 'Sepal.Width' will be left
iris %>% except(from("Petal.Length"))

# except first column
iris %n_d% items(1)

# if_val examples
# From SPSS: RECODE QVAR(1 THRU 5=1)(6 THRU 10=2)(11 THRU HI=3)(ELSE=0).
set.seed(123)
qvar = sample((-5):20, 50, replace = TRUE)
recode(qvar, 1 %thru% 5 ~ 1, 6 %thru% 10 ~ 2, 11 %thru% hi ~ 3, other ~ 0)
# the same result
recode(qvar, 1 %thru% 5 ~ 1, 6 %thru% 10 ~ 2, ge(11) ~ 3, other ~ 0)
```

`cro` *Cross tabulation with support of labels, weights and multiple response variables.*

Description

- `cro`, `cro_cases` build a contingency table of the counts.
- `cro_cpct`, `cro_cpct_responses` build a contingency table of the column percent. These functions give different results only for multiple response variables. For `cro_cpct` base of percent is number of valid cases. Case is considered as valid if it has at least one non-NA value. So for multiple response variables sum of percent may be greater than 100. For `cro_cpct_responses` base of percent is number of valid responses. Multiple response variables can have several responses for single case. Sum of percent of `cro_cpct_responses` always equals to 100%.
- `cro_rpct` build a contingency table of the row percent. Base for percent is number of valid cases.
- `cro_tpct` build a contingency table of the table percent. Base for percent is number of valid cases.
- `calc_cro_*` are the same as above but evaluate their arguments in the context of the first argument `data`.

- `total` auxiliary function - creates variables with 1 for valid case of its argument `x` and NA in opposite case.

You can combine tables with add_rows and merge.etable. For sorting table see tab_sort_asc. To provide multiple-response variables as arguments use mrset for multiples with category encoding and mdset for multiples with dichotomy (dummy) encoding. To compute statistics with nested variables/banners use nest. For more sophisticated interface with modern piping via `magrittr` see tables.

Usage

```
cro(cell_vars, col_vars = total(), row_vars = NULL, weight = NULL,
  subgroup = NULL, total_label = NULL, total_statistic = "u_cases",
  total_row_position = c("below", "above", "none"))

cro_cases(cell_vars, col_vars = total(), row_vars = NULL, weight = NULL,
  subgroup = NULL, total_label = NULL, total_statistic = "u_cases",
  total_row_position = c("below", "above", "none"))

cro_cpct(cell_vars, col_vars = total(), row_vars = NULL, weight = NULL,
  subgroup = NULL, total_label = NULL, total_statistic = "u_cases",
  total_row_position = c("below", "above", "none"))

cro_rpct(cell_vars, col_vars = total(), row_vars = NULL, weight = NULL,
  subgroup = NULL, total_label = NULL, total_statistic = "u_cases",
  total_row_position = c("below", "above", "none"))

cro_tpct(cell_vars, col_vars = total(), row_vars = NULL, weight = NULL,
  subgroup = NULL, total_label = NULL, total_statistic = "u_cases",
  total_row_position = c("below", "above", "none"))

cro_cpct_responses(cell_vars, col_vars = total(), row_vars = NULL,
  weight = NULL, subgroup = NULL, total_label = NULL,
  total_statistic = "u_responses", total_row_position = c("below", "above",
  "none"))

calc_cro(data, cell_vars, col_vars = total(), row_vars = NULL,
  weight = NULL, subgroup = NULL, total_label = NULL,
  total_statistic = "u_cases", total_row_position = c("below", "above",
  "none"))

calc_cro_cases(data, cell_vars, col_vars = total(), row_vars = NULL,
  weight = NULL, subgroup = NULL, total_label = NULL,
  total_statistic = "u_cases", total_row_position = c("below", "above",
  "none"))

calc_cro_cpct(data, cell_vars, col_vars = total(), row_vars = NULL,
  weight = NULL, subgroup = NULL, total_label = NULL,
  total_statistic = "u_cases", total_row_position = c("below", "above",
```

```
  "none"))

calc_cro_rpct(data, cell_vars, col_vars = total(), row_vars = NULL,
  weight = NULL, subgroup = NULL, total_label = NULL,
  total_statistic = "u_cases", total_row_position = c("below", "above",
  "none"))

calc_cro_tpct(data, cell_vars, col_vars = total(), row_vars = NULL,
  weight = NULL, subgroup = NULL, total_label = NULL,
  total_statistic = "u_cases", total_row_position = c("below", "above",
  "none"))

calc_cro_cpct_responses(data, cell_vars, col_vars = total(),
  row_vars = NULL, weight = NULL, subgroup = NULL, total_label = NULL,
  total_statistic = "u_responses", total_row_position = c("below", "above",
  "none"))

total(x = 1, label = "#Total")
```

Arguments

cell_vars	vector/data.frame/list. Variables on which percentage/cases will be computed. Use mrset/mdset for multiple-response variables.
col_vars	vector/data.frame/list. Variables which breaks table by columns. Use mrset/mdset for multiple-response variables.
row_vars	vector/data.frame/list. Variables which breaks table by rows. Use mrset/mdset for multiple-response variables.
weight	numeric vector. Optional cases weights. Cases with NA's, negative and zero weights are removed before calculations.
subgroup	logical vector. You can specify subgroup on which table will be computed.
total_label	By default "#Total". You can provide several names - each name for each total statistics.
total_statistic	By default it is "u_cases" (unweighted cases). Possible values are "u_cases", "u_responses", "u_cpct", "u_rpct", "u_tpct", "w_cases", "w_responses", "w_cpct", "w_rpct", "w_tpct". "u_" means unweighted statistics and "w_" means weighted statistics.
total_row_position	Position of total row in the resulting table. Can be one of "below", "above", "none".
data	data.frame in which context all other arguments will be evaluated (for calc_cro_*).
x	vector/data.frame of class 'category'/'dichotomy'.
label	character. Label for total variable.

Value

object of class 'etable'. Basically it's a data.frame but class is needed for custom methods.

See Also

tables, fre, cro_fun.

Examples

```
data(mtcars)
mtcars = apply_labels(mtcars,
                      mpg = "Miles/(US) gallon",
                      cyl = "Number of cylinders",
                      disp = "Displacement (cu.in.)",
                      hp = "Gross horsepower",
                      drat = "Rear axle ratio",
                      wt = "Weight (1000 lbs)",
                      qsec = "1/4 mile time",
                      vs = "Engine",
                      vs = c("V-engine" = 0,
                             "Straight engine" = 1),
                      am = "Transmission",
                      am = c("Automatic" = 0,
                             "Manual"=1),
                      gear = "Number of forward gears",
                      carb = "Number of carburetors"
)

calculate(mtcars, cro(am, vs))
calc_cro(mtcars, am, vs) # the same result

# column percent with multiple banners
calculate(mtcars, cro_cpct(cyl, list(total(), vs, am)))
calc_cro_cpct(mtcars, cyl, list(total(), vs, am)) # the same result

# nested banner
calculate(mtcars, cro_cpct(cyl, list(total(), vs %nest% am)))

# stacked variables
calculate(mtcars, cro(list(cyl, carb), list(total(), vs %nest% am)))

# nested variables
calculate(mtcars, cro_cpct(am %nest% cyl, list(total(), vs)))

# row variables
calculate(mtcars, cro_cpct(cyl, list(total(), vs), row_vars = am))

# several totals above table
calculate(mtcars, cro_cpct(cyl,
              list(total(), vs),
              row_vars = am,
              total_row_position = "above",
              total_label = c("number of cases", "row %"),
              total_statistic = c("u_cases", "u_rpct")
              ))
```

```
# multiple-choice variable
# brands - multiple response question
# Which brands do you use during last three months?
set.seed(123)
brands = data.frame(t(replicate(20,sample(c(1:5,NA),4,replace = FALSE))))
# score - evaluation of tested product
score = sample(-1:1,20,replace = TRUE)
var_lab(brands) = "Used brands"
val_lab(brands) = make_labels("
                              1 Brand A
                              2 Brand B
                              3 Brand C
                              4 Brand D
                              5 Brand E
                              ")

var_lab(score) = "Evaluation of tested brand"
val_lab(score) = num_lab("
                            -1 Dislike it
                            0 So-so
                            1 Like it
                            ")

cro_cpct(mrset(brands), list(total(), score))
# responses
cro_cpct_responses(mrset(brands), list(total(), score))
```

`cro_fun` *Cross-tabulation with custom summary function.*

Description

- `cro_mean`, `cro_sum`, `cro_median` calculate mean/sum/median by groups. NA's are always omitted.
- `cro_mean_sd_n` calculates mean, standard deviation and N simultaneously. Mainly intended for usage with significance_means.
- `cro_pearson`, `cro_spearman` calculate correlation of first variable in each data.frame in `cell_vars` with other variables. NA's are removed pairwise.
- `cro_fun`, `cro_fun_df` return table with custom summary statistics defined by `fun` argument. NA's treatment depends on your `fun` behavior. To use weight you should have formal `weight` argument in `fun` and some logic for its processing inside. Several functions with weight support are provided - see w_mean. `cro_fun` applies `fun` on each variable in `cell_vars` separately, `cro_fun_df` gives to `fun` each data.frame in `cell_vars` as a whole. So `cro_fun(iris[, -5], iris$Species, fun = mean)` gives the same result as `cro_fun_df(iris[, -5], iris$Spec` For `cro_fun_df` names of `cell_vars` will converted to labels if they are available before the `fun` will be applied. Generally it is recommended that `fun` will always return object of the same form. Row names/vector names of `fun` result will appear in the row labels of the table and column names/names of list will appear in the column labels. If your `fun` returns

data.frame/matrix/list with element named 'row_labels' then this element will be used as row labels. And it will have precedence over rownames.

- `calc_cro_*` are the same as above but evaluate their arguments in the context of the first argument `data`.
- `combine_functions` is auxiliary function for combining several functions into one function for usage with `cro_fun`/`cro_fun_df`. Names of arguments will be used as statistic labels. By default, results of each function are combined with `c`. But you can provide your own method function with `method` argument. It will be applied as in the expression `do.call(method, list_of_functions_results)`. Particular useful method is `list`. When it used then statistic labels will appear in the column labels. See examples. Also you may be interested in `data.frame`, `rbind`, `cbind` methods.

Usage

```
cro_fun(cell_vars, col_vars = total(), row_vars = total(label = ""),
  weight = NULL, subgroup = NULL, fun, ..., unsafe = FALSE)

cro_fun_df(cell_vars, col_vars = total(), row_vars = total(label = ""),
  weight = NULL, subgroup = NULL, fun, ..., unsafe = FALSE)

cro_mean(cell_vars, col_vars = total(), row_vars = total(label = ""),
  weight = NULL, subgroup = NULL)

cro_mean_sd_n(cell_vars, col_vars = total(), row_vars = total(label = ""),
  weight = NULL, subgroup = NULL, weighted_valid_n = FALSE,
  labels = NULL)

cro_sum(cell_vars, col_vars = total(), row_vars = total(label = ""),
  weight = NULL, subgroup = NULL)

cro_median(cell_vars, col_vars = total(), row_vars = total(label = ""),
  weight = NULL, subgroup = NULL)

cro_pearson(cell_vars, col_vars = total(), row_vars = total(label = ""),
  weight = NULL, subgroup = NULL)

cro_spearman(cell_vars, col_vars = total(), row_vars = total(label = ""),
  weight = NULL, subgroup = NULL)

calc_cro_fun(data, cell_vars, col_vars = total(), row_vars = total(label =
  ""), weight = NULL, subgroup = NULL, fun, ..., unsafe = FALSE)

calc_cro_fun_df(data, cell_vars, col_vars = total(), row_vars = total(label
  = ""), weight = NULL, subgroup = NULL, fun, ..., unsafe = FALSE)

calc_cro_mean(data, cell_vars, col_vars = total(), row_vars = total(label =
  ""), weight = NULL, subgroup = NULL)
```

```
calc_cro_mean_sd_n(data, cell_vars, col_vars = total(),
  row_vars = total(label = ""), weight = NULL, subgroup = NULL,
  weighted_valid_n = FALSE, labels = NULL)

calc_cro_sum(data, cell_vars, col_vars = total(), row_vars = total(label =
  ""), weight = NULL, subgroup = NULL)

calc_cro_median(data, cell_vars, col_vars = total(), row_vars = total(label
  = ""), weight = NULL, subgroup = NULL)

calc_cro_pearson(data, cell_vars, col_vars = total(), row_vars = total(label
  = ""), weight = NULL, subgroup = NULL)

calc_cro_spearman(data, cell_vars, col_vars = total(),
  row_vars = total(label = ""), weight = NULL, subgroup = NULL)

combine_functions(..., method = c)
```

Arguments

cell_vars	vector/data.frame/list. Variables on which summary function will be computed.
col_vars	vector/data.frame/list. Variables which breaks table by columns. Use mrset/mdset for multiple-response variables.
row_vars	vector/data.frame/list. Variables which breaks table by rows. Use mrset/mdset for multiple-response variables.
weight	numeric vector. Optional cases weights. Cases with NA's, negative and zero weights are removed before calculations.
subgroup	logical vector. You can specify subgroup on which table will be computed.
fun	custom summary function. Generally it is recommended that fun will always return object of the same form. Rownames/vector names of fun result will appear in the row labels of the table and column names/names of list will appear in the column labels. To use weight you should have formal weight argument in fun and some logic for its processing inside. For cro_fun_df fun will receive data.table with all names converted to variable labels (if labels exists). So it is not recommended to rely on original variables names in your fun.
...	further arguments for fun in cro_fun/cro_fun_df or functions for combine_functions. Ignored in cro_fun/cro_fun_df if unsafe is TRUE.
unsafe	logical/character If not FALSE than fun will be evaluated as is. It can lead to significant increase in the performance. But there are some limitations. For cro_fun it means that your function fun should return vector. If length of this vector is greater than one than you should provide with unsafe argument vector of unique labels for each element of this vector. There will be no attempts to automatically make labels for the results of fun. For cro_fun_df your function should return vector or list/data.frame (optionally with 'row_labels' element - statistic labels). If unsafe is TRUE or not logical then further arguments (...) for fun will be ignored.

weighted_valid_n
: logical. Should we show weighted valid N in cro_mean_sd_n? By default it is FALSE.

labels
: character vector of length 3. Labels for mean, standard deviation and valid N in cro_mean_sd_n.

data
: data.frame in which context all other arguments will be evaluated (for calc_cro_*).

method
: function which will combine results of multiple functions in combine_functions. It will be applied as in the expression do.call(method, list_of_functions_results). By default it is c.

Value

object of class 'etable'. Basically it's a data.frame but class is needed for custom methods.

See Also

tables, fre, cro.

Examples

```
data(mtcars)
mtcars = apply_labels(mtcars,
                      mpg = "Miles/(US) gallon",
                      cyl = "Number of cylinders",
                      disp = "Displacement (cu.in.)",
                      hp = "Gross horsepower",
                      drat = "Rear axle ratio",
                      wt = "Weight (1000 lbs)",
                      qsec = "1/4 mile time",
                      vs = "Engine",
                      vs = c("V-engine" = 0,
                             "Straight engine" = 1),
                      am = "Transmission",
                      am = c("Automatic" = 0,
                             "Manual"=1),
                      gear = "Number of forward gears",
                      carb = "Number of carburetors"
)


# Simple example - there is special shortcut for it - 'cro_mean'
calculate(mtcars, cro_fun(list(mpg, disp, hp, wt, qsec),
                                col_vars = list(total(), am),
                                row_vars = vs,
                                fun = mean)
)

# the same result
calc_cro_fun(mtcars, list(mpg, disp, hp, wt, qsec),
                     col_vars = list(total(), am),
                     row_vars = vs,
```

```
                    fun = mean
)

# The same example with 'subgroup'
calculate(mtcars, cro_fun(list(mpg, disp, hp, wt, qsec),
                              col_vars = list(total(), am),
                              row_vars = vs,
                              subgroup = vs == 0,
                              fun = mean)
)

# 'combine_functions' usage
calculate(mtcars, cro_fun(list(mpg, disp, hp, wt, qsec),
                          col_vars = list(total(), am),
                          row_vars = vs,
                          fun = combine_functions(Mean = mean,
                                                  'Std. dev.' = sd,
                                                  'Valid N' = valid_n)
))
# 'combine_functions' usage - statistic labels in columns
calculate(mtcars, cro_fun(list(mpg, disp, hp, wt, qsec),
                          col_vars = list(total(), am),
                          row_vars = vs,
                          fun = combine_functions(Mean = mean,
                                                  'Std. dev.' = sd,
                                                  'Valid N' = valid_n,
                                                  method = list
                                                  )
))

# 'summary' function
calculate(mtcars, cro_fun(list(mpg, disp, hp, wt, qsec),
                          col_vars = list(total(), am),
                          row_vars = list(total(), vs),
                          fun = summary
))

# comparison 'cro_fun' and 'cro_fun_df'
calculate(mtcars, cro_fun(
                        sheet(mpg, disp, hp, wt, qsec),
                        col_vars = am,
                        fun = mean
                        )
)

# same result
calculate(mtcars, cro_fun_df(
                        sheet(mpg, disp, hp, wt, qsec),
                        col_vars = am,
                        fun = colMeans
                        )
)
```

```
# usage for 'cro_fun_df' which is not possible for 'cro_fun'
# linear regression by groups
calculate(mtcars, cro_fun_df(
                      sheet(mpg, disp, hp, wt, qsec),
                      col_vars = am,
                      fun = function(x){
                            frm = reformulate(".", response = names(x)[1])
                            model = lm(frm, data = x)
                            sheet(
                                'Coef. estimate' = coef(model),
                                 confint(model)
                                 )
                      }
))
```

`default_dataset` *Get or set reference to default dataset. Experimental feature.*

Description

Use data.frame or data.frame name to set it as default. Use NULL as an argument to disable default dataset. If argument is missing then function will return reference to default dataset. Use ref to modify it. Also see .compute for usage patterns.

Usage

```
default_dataset(x)
```

Arguments

x	data.frame or data.frame name which we want to make default for some operations.

Value

formula reference to default dataset or NULL

See Also

ref

Examples

```
data(iris)
default_iris = iris
default_dataset(default_iris) # set default dataset

.compute({
```

```
    new_col = 1
    Sepal.Length = Sepal.Length*2
})

# for comparison

iris$new_col = 1
iris$Sepal.Length = iris$Sepal.Length*2
identical(iris, default_iris) # should be TRUE

default_dataset(NULL) # disable default dataset
```

`do_repeat` *Repeats the same transformations on a specified set of variables/values*

Description

Repeats the same transformations on a specified set of variables/values

Usage

```
do_repeat(data, ...)

.do_repeat(...)

as_is(...)
```

Arguments

`data`	data.frame/list. If `data` is list then `do_repeat` will be applied to each element of the list.
`...`	stand-in name(s) followed by equals sign and a vector/list of replacement variables or values. They can be numeric/characters or variables names. Names at the top-level can be unquoted (non-standard evaluation). Quoted characters also considered as variables names. To avoid this behavior use `as_is` function. For standard evaluation of parameters you can surround them by round brackets. Also you can use %to% operator and other criteria functions. Last argument should be expression in curly brackets which will be evaluated in the scope of data.frame `data`. See examples.

Details

There is a special constant `.N` which equals to number of cases in `data` for usage in expression inside `do_repeat`. Also there are a variables `.item_num` which is equal to the current iteration number and `.item_value` which is named list with current stand-in variables values.

Value

transformed data.frame `data`

See Also

compute, do_if

Examples

```
data(iris)
scaled_iris = do_repeat(iris,
                        i = Sepal.Length %to% Petal.Width,
                        {
                            i = scale(i)
                        })
head(scaled_iris)

# several stand-in names and standard evaluattion
old_names = qc(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)
new_names = paste0("scaled_", old_names)
scaled_iris = do_repeat(iris,
                        orig = ((old_names)),
                        scaled = ((new_names)),
                        {
                            scaled = scale(orig)
                        })
head(scaled_iris)

# numerics
new_df = data.frame(id = 1:20)
# note the automatic creation of the sequence of variables
new_df = do_repeat(new_df,
                   item = i1 %to% i3,
                   value = c(1, 2, 3),
                   {
                       item = value
                   })
head(new_df)

# the same result with internal variable '.item_num'
new_df = data.frame(id = 1:20)
new_df = do_repeat(new_df,
                   item = i1 %to% i3,
                   {
                       item = .item_num
                   })
head(new_df)

# functions
set.seed(123)
new_df = data.frame(id = 1:20)
new_df = do_repeat(new_df,
```

```
                    item = c(i1, i2, i3),
                    fun = c("rnorm", "runif", "rexp"),
                    {
                        item = fun(.N)
                    })
head(new_df)
```

`drop_empty_rows` *Drop empty (with all NA's) rows/columns from data.frame/table*

Description

By default tables produced by functions tables, cro, cro_fun and cro_fun_df are created with all possible value labels. If values for this labels are absent in variable there are NA's in rows and columns. `drop_empty_rows`/`drop_empty_columns` are intended to remove these empty rows/columns. `drop_r` and `drop_c` are the same functions with shorter names. `drop_rc` drops rows and columns simultaneously.

Usage

```
drop_empty_rows(x, excluded_rows = NULL, excluded_columns = NULL)

drop_empty_columns(x, excluded_rows = NULL, excluded_columns = NULL)

drop_r(x, excluded_rows = NULL, excluded_columns = NULL)

drop_c(x, excluded_rows = NULL, excluded_columns = NULL)

drop_rc(x)
```

Arguments

`x`	data.frame/etable(result of cro and etc.)
`excluded_rows`	character/logical/numeric rows which won't be dropped and in which NAs won't be counted. If it is characters then they will be considered as pattern/vector of patterns. Patterns will be matched with Perl-style regular expression with values in the first column of `x` (see grep, `perl = TRUE` argument). Rows which have such patterns will be excluded. By default for class 'etable' pattern is "#" because "#" marks totals in the result of cro.
`excluded_columns`	logical/numeric/characters columns which won't be dropped and in which NAs won't be counted. By default for class 'etable' it is first column - column with labels in table.

Value

data.frame with removed rows/columns

Examples

```
data(mtcars)
mtcars = apply_labels(mtcars,
            vs = "Engine",
            vs = num_lab("
                      0 V-engine
                      1 Straight engine
                      9 Other
                      "),
            am = "Transmission",
            am = num_lab("
                     0 Automatic
                     1 Manual
                     9 Other
                     ")
         )
with_empty = calculate(mtcars, cro(am, vs))

drop_empty_rows(with_empty)
drop_empty_columns(with_empty)
drop_rc(with_empty)
```

`experimental` *Experimental functions for operations with default dataset*

Description

Workflow for these functions is rather simple. You should set up default data.frame with default_dataset and then operate with it without any reference to your data.frame. There are two kinds of operations. The first kind modify default dataset, the second kind will be evaluated in the context of the default dataset but doesn't modify it. It is not recommended to use one of these functions in the scope of another of these functions. By now their performance is not so high, especially `.do_if`/`.modify_if` can be very slow.

Usage

```
.modify(expr)

.modify_if(cond, expr)

.do_if(cond, expr)

.compute(expr)

.calculate(expr, use_labels = FALSE)

.calc(expr, use_labels = FALSE)
```

```
.val_lab(...)

.var_lab(...)

.if_val(x, ...)

.recode(x, ...)

.fre(...)

.cro(...)

.cro_cpct(...)

.cro_rpct(...)

.cro_tpct(...)

.cro_mean(...)

.cro_sum(...)

.cro_median(...)

.cro_fun(...)

.cro_fun_df(...)
```

Arguments

`expr`	set of expressions in curly brackets which will be evaluated in the context of default dataset
`cond`	logical vector/expression
`use_labels`	logical. Experimental feature. If it equals to `TRUE` then we will try to replace variable names with labels. Many base R functions which show variable names will show labels.
`...`	further arguments
`x`	vector/data.frame - variable names in the scope of default dataset

Details

Functions which modify default dataset:

- `.modify` Add and modify variables inside default data.frame. See modify.
- `.compute` Shortcut for `.modify`. Name is inspired by SPSS COMPUTE operator. See modify.
- `.modify_if` Add and modify variables inside subset of default data.frame. See modify_if.

- `.do_if` Shortcut for `.modify_if`. Name is inspired by SPSS DO IF operator. See modify_if.
- `.where` Leave subset of default data.frame which meet condition. See where, subset.
- `.recode` Change, rearrange or consolidate the values of an existing variable inside default data.frame. See recode.
- `.if_val` Shortcut for `.recode`. See recode.

Other functions:

- `.var_lab` Return variable label from default dataset. See var_lab.
- `.val_lab` Return value labels from default dataset. See val_lab.
- `.fre` Simple frequencies of variable in the default data.frame. See fre.
- `.cro`/`.cro_cpct`/`.cro_rpct`/`.cro_tpct` Simple crosstabulations of variable in the default data.frame. See cro.
- `.cro_mean`/`.cro_sum`/`.cro_median`/`.cro_fun`/`.cro_fun_df` Simple crosstabulations of variable in the default data.frame. See cro_fun.
- `.calculate` Evaluate arbitrary expression in the context of data.frame. See calculate.

Examples

```
data(mtcars)

default_dataset(mtcars) # set mtcars as default dataset

# calculate new variables
.compute({
    mpg_by_am = ave(mpg, am, FUN = mean)
    hi_low_mpg = ifs(mpg<mean(mpg) ~ 0, TRUE ~ 1)
})

# set labels
.apply_labels(
    mpg = "Miles/(US) gallon",
    cyl = "Number of cylinders",
    disp = "Displacement (cu.in.)",
    hp = "Gross horsepower",
    mpg_by_am = "Average mpg for transimission type",
    hi_low_mpg = "Miles per gallon",
    hi_low_mpg = num_lab("
                    0 Low
                    1 High
                    "),

    vs = "Engine",
    vs = num_lab("
                    0 V-engine
                    1 Straight engine
                "),

    am = "Transmission",
    am = num_lab("
```

```
                    0 Automatic
                    1 Manual
                         ")
)
# calculate frequencies
.fre(hi_low_mpg)
.cro(cyl, hi_low_mpg)
.cro_mean(data.frame(mpg, disp, hp), vs)

# disable default dataset
default_dataset(NULL)

# Example of .recode

data(iris)

default_dataset(iris) # set iris as default dataset

.recode(Sepal.Length, lo %thru% median(Sepal.Length) ~ "small", other ~ "large")

.fre(Sepal.Length)

# example of .do_if

.do_if(Species == "setosa",{
     Petal.Length = NA
     Petal.Width = NA
})

.cro_mean(data.frame(Petal.Length, Petal.Width), Species)

# disable default dataset
default_dataset(NULL)
```

`expss` *expss: Tables with Labels and Some Useful Functions from Spreadsheets and SPSS Statistics*

Description

'expss' package implements some popular functions from spreadsheets and SPSS Statistics software. Implementations are not complete copies of their originals. I try to make them consistent with other R functions. See examples in the vignette and in the help.

Excel

- IF ifelse
- AVERAGE mean_row
- SUM sum_row

- MIN min_row
- MAX max_row
- VLOOKUP vlookup
- COUNTIF count_if
- AVERAGEIF mean_row_if
- SUMIF sum_row_if
- MINIF min_row_if
- MAXIF max_row_if
- IFS ifs
- IFNA if_na
- MATCH match_row
- INDEX index_row
- PIVOT TABLES tables, cro_fun, cro

SPSS

- COMPUTE compute
- RECODE recode
- COUNT count_row_if
- DO IF do_if
- DO REPEAT do_repeat
- VARIABLE LABELS var_lab
- VALUE LABELS val_lab
- ANY any_in_row
- FREQUENCIES fre
- CROSSTABS cro
- CUSTOM TABLES tables

`expss.options` *Options for contolling behaviour of the package*

Description

All options can be set with `options(option.name = option.value)` or with special functions (see below). You can get value of option with `getOption("option.name")`.

- `expss.digits` Number of digits after decimal separator which will be shown for tables. This parameter is supported in the datatable, htmlTable.etable and `print` methods. `NULL` is default and means one digit. `NA` means no rounding. There is convenience function for this option: `expss_digits`.

- `expss.enable_value_labels_support` By default, all labelled variables will use labels as labels for factor levels when factor is called. So any function which calls factor/as.factor will use value labels. In details this option changes behavior of two methods for class `labelled` - `as.character` and `unique` - on which `factor` depends entirely. If you have compatibility problems set this option to zero: `options(expss.enable_value_labels_support = 0)`. There are shortcuts for these options: `expss_enable_value_labels_support()` and `expss_disable_value_labels_`
- `expss.output` By default tables are printed in the console. You can change this behavior by setting this option. There are four possible values: `'rnotebook'`, `'viewer'`, `'commented'` or `'raw'`. First option is useful when you run your code in the R Notebook - output will be rendered to nice HTML. The second option will render tables to RStudio viewer. `knitr` is supported automatically via `knit_print` method. `'commented'` prints default output to the console with comment symbol (#) at the beginning of the each line. With comment symbol you can easily copy and paste your output into the script. Option `raw` disables any formatting and all tables are printed as data.frames. Shortcuts for options: `expss_output_default()`, `expss_output_raw()`, `expss_output_viewer()`, `expss_output_commented()` and `expss_output_rnotebook()`.

Usage

```
expss_digits(digits = NULL)

get_expss_digits()

expss_enable_value_labels_support()

expss_disable_value_labels_support()

expss_output_default()

expss_output_commented()

expss_output_raw()

expss_output_viewer()

expss_output_rnotebook()
```

Arguments

`digits`	integer. Number of digits after decimal point. `NULL` is default and means 1 digit. `NA` means no rounding.

`fctr` *Convert labelled variable to factor*

Description

fctr converts variable to factor. It force labels usage as factor labels for labelled variables even if 'expss.enable_value_labels_support' set to 0. For other types of variables base factor is called. Factor levels are constructed as values labels. If label doesn't exist for particular value then this value remain as is - so there is no information lost. This levels look like as "Variable_label|Value label" if argument prepend set to TRUE.

Usage

```
fctr(x, ..., drop_unused_labels = FALSE, prepend_var_lab = TRUE)
```

Arguments

x	a vector of data with labels.
...	optional arguments for factor
drop_unused_labels	logical. Should we drop unused value labels? Default is FALSE.
prepend_var_lab	logical. Should we prepend variable label before value labels? Default is TRUE.

Value

an object of class factor. For details see base factor documentation.

See Also

values2labels, names2labels, val_lab, var_lab. Materials for base functions: factor, as.factor, ordered, as.ordered

Examples

```
data(mtcars)

var_lab(mtcars$am) = "Transmission"
val_lab(mtcars$am) = c(automatic = 0, manual=1)

summary(lm(mpg ~ am, data = mtcars)) # no labels
summary(lm(mpg ~ fctr(am), data = mtcars)) # with labels
summary(lm(mpg ~ fctr(unvr(am)), data = mtcars)) # without variable label
```

fre *Simple frequencies with support of labels, weights and multiple response variables.*

Description

fre returns data.frame with six columns: labels or values, counts, valid percent (excluding NA), percent (with NA), percent of responses(for single-column x it equals to valid percent) and cumulative percent of responses.

Usage

```
fre(x, weight = NULL, drop_unused_labels = TRUE, prepend_var_lab = FALSE)
```

Arguments

x	vector/data.frame/list. data.frames are considered as multiple response variables. If x is list then vertically stacked frequencies for each element of list will be generated,
weight	numeric vector. Optional case weights. NA's and negative weights treated as zero weights.
drop_unused_labels	logical. Should we drop unused value labels? Default is TRUE.
prepend_var_lab	logical. Should we prepend variable label before value labels? By default we will add variable labels to value labels only if x or predictor is list (several variables).

Value

object of class 'etable'. Basically it's a data.frame but class is needed for custom methods.

Examples

```
data(mtcars)
mtcars = modify(mtcars,{
    var_lab(vs) = "Engine"
    val_lab(vs) = c("V-engine" = 0,
                    "Straight engine" = 1)
    var_lab(am) = "Transmission"
    val_lab(am) = c(automatic = 0,
                    manual=1)
})

fre(mtcars$vs)

# stacked frequencies
fre(list(mtcars$vs, mtcars$am))
```

```
# multiple-choice variable
# brands - multiple response question
# Which brands do you use during last three months?
set.seed(123)
brands = data.frame(t(replicate(20,sample(c(1:5,NA),4,replace = FALSE))))
# score - evaluation of tested product
score = sample(-1:1,20,replace = TRUE)
var_lab(brands) = "Used brands"
val_lab(brands) = make_labels("
                              1 Brand A
                              2 Brand B
                              3 Brand C
                              4 Brand D
                              5 Brand E
                              ")

var_lab(score) = "Evaluation of tested brand"
val_lab(score) = make_labels("
                             -1 Dislike it
                             0 So-so
                             1 Like it
                             ")

fre(brands)

# stacked frequencies
fre(list(score, brands))
```

`from_text` *Make data.frame from text*

Description

Convert delimited text lines to data.frame. Blank lines are always skipped, trailing whitespaces are trimmed. You can use comments with '#' inside your text. For details see read.table.

Usage

```
from_text(text, header = TRUE, sep = "", quote = "", dec = ".",
  encoding = "unknown", ...)

from_text_csv(text, header = TRUE, sep = ",", quote = "", dec = ".",
  encoding = "unknown", ...)

from_text_csv2(text, header = TRUE, sep = ";", quote = "", dec = ",",
  encoding = "unknown", ...)
```

```
from_text_tab(text, header = TRUE, sep = "\t", quote = "", dec = ".",
  encoding = "unknown", ...)

from_text_tab2(text, header = TRUE, sep = "\t", quote = "", dec = ",",
  encoding = "unknown", ...)
```

Arguments

text	character/vector of characters
header	a logical value indicating whether the text contains the names of the variables as its first line.
sep	the field separator character. Values on each line of the file are separated by this character. If sep = "" (the default for read.table) the separator is 'white space', that is one or more spaces, tabs, newlines or carriage returns.
quote	the set of quoting characters. To disable quoting altogether, use quote = "".
dec	the character used in the file for decimal points.
encoding	encoding to be assumed for input strings. It is used to mark character strings as known to be in Latin-1 or UTF-8 (see read.table).
...	further parameters which will be passed to read.table.

Value

data.frame

Examples

```
from_text("
# simple data.frame
     a b   c
     1 2.5 a
     4 5.5 b
     7 8.5 c
")
```

---

htmlTable.etable *Outputting HTML tables in RStudio viewer/R Notebooks*

Description

This is method for rendering results of fre/cro/tables in Shiny/RMarkdown/Jupyter notebooks and etc. For detailed description of function and its arguments see htmlTable. You may be interested in `expss_output_viewer()` for automatical rendering tables in the RStudio viewer or `expss_output_rnotebook()` for rendering in the R notebooks. See expss.options. `repr_html` is method for rendering table in the Jupyter notebooks and `knit_print` is method for rendering table in the `knitr` HTML-documents. Jupyter notebooks and `knitr` documents are supported automatically but in the R notebooks it is needed to set output to notebook via `expss_output_rnotebook()`.

Usage

```
## S3 method for class 'etable'
htmlTable(x, digits = get_expss_digits(),
  escape.html = FALSE, ..., row_groups = TRUE)

knit_print.etable(x, digits = get_expss_digits(), escape.html = FALSE, ...)

repr_html.etable(obj, digits = get_expss_digits(), escape.html = FALSE, ...)

repr_text.etable(obj, digits = get_expss_digits(), ...)
```

Arguments

x	a data object of class 'etable' - result of fre/cro and etc.
digits	integer By default, all numeric columns are rounded to one digit after decimal separator. Also you can set this argument by setting option 'expss.digits' - for example, `expss_digits(2)`. If it is NA than all numeric columns remain unrounded.
escape.html	logical: should HTML characters be escaped? Defaults to FALSE.
...	further parameters for htmlTable.
row_groups	logical Should we create row groups? TRUE by default.
obj	a data object of class 'etable' - result of fre/cro and etc.

Value

Returns a string of class htmlTable

Examples

```
## Not run:
data(mtcars)
mtcars = apply_labels(mtcars,
                      mpg = "Miles/(US) gallon",
                      cyl = "Number of cylinders",
                      disp = "Displacement (cu.in.)",
                      hp = "Gross horsepower",
                      drat = "Rear axle ratio",
                      wt = "Weight (1000 lbs)",
                      qsec = "1/4 mile time",
                      vs = "Engine",
                      vs = c("V-engine" = 0,
                             "Straight engine" = 1),
                      am = "Transmission",
                      am = c("Automatic" = 0,
                             "Manual"=1),
                      gear = "Number of forward gears",
                      carb = "Number of carburetors"
)
```

```
expss_output_viewer()
mtcars %>%
     tab_cols(total(), am %nest% vs) %>%
     tab_cells(mpg, hp) %>%
     tab_stat_mean() %>%
     tab_cells(cyl) %>%
     tab_stat_cpct() %>%
     tab_pivot()

expss_output_default()


## End(Not run)
```

`if_na` *Replace NA values in vector/data.frame/matrix/list with supplied value*

Description

Function replaces NA values in vector/data.frame/matrix/list with supplied value. If x is vector then `if_na(x) = 99` is equivalent to `x[is.na(x)] = 99`. In more complex cases when x is data.frame/matrix/list this function tries to replace NA recursively. If replacement value is vector/data.frame/matrix/list then `if_na` uses for replacement values from appropriate places. For example if both `x` and `value` are vectors then `if_na(x) = value` is equivalent to `x[is.na(x)] = value[is.na(x)]`. Single column/row value recycled to conform to `x`. See examples.

Usage

```
if_na(x, value)

if_na(x) <- value

x %if_na% value
```

Arguments

`x`	vector/matrix/data.frame/list
`value`	vector/matrix/data.frame/list

Value

x with replaced NA

See Also

For reverse operation see na_if.

Examples

```
# simple case
a = c(NA, 2, 3, 4, NA)
if_na(a, 99)

# the same result
a %if_na% 99

# the same result
if_na(a) = 99
a # c(99, 2, 3, 4, 99)


# replacement with values from other variable
a = c(NA, 2, 3, 4, NA)
if_na(a) = 1:5
a # 1:5

# replacement with group means

# make data.frame
set.seed(123)
group = sample(1:3, 30, replace = TRUE)
param = runif(30)
param[sample(30, 10)] = NA # place 10 NA's
df = data.frame(group, param)

# replace NA's with group means
df = compute(df, {
        if_na(param) = window_fun(param, group, mean_col)
    })

df

# replacement with column means

# make data.frame
set.seed(123)
x1 = runif(30)
x2 = runif(30)
x3 = runif(30)
x1[sample(30, 10)] = NA # place 10 NA's
x2[sample(30, 10)] = NA # place 10 NA's
x3[sample(30, 10)] = NA # place 10 NA's

df = data.frame(x1, x2, x3)

# replace NA's with column means
if_na(df) = t(mean_col(df))

df
```

`if_val` *Change, rearrange or consolidate the values of an existing/new variable. Inspired by RECODE command from SPSS.*

Description

`recode` change, rearrange or consolidate the values of an existing variable based on conditions. Design of this function inspired by RECODE from SPSS. Sequence of recodings provided in the form of formulas. For example, 1:2 ~ 1 means that all 1's and 2's will be replaced with 1. Each value will recoded only once. In the assignment form `recode(...) = ...` of this function values which doesn't meet any condition remain unchanged. In case of the usual form `... = recode(...)` values which doesn't meet any condition will be replaced with NA. One can use values or more sophisticated logical conditions and functions as a condition. There are several special functions for usage as criteria - for details see criteria. Simple common usage looks like: `recode(x, 1:2 ~ -1, 3 ~ 0, 1:2 ~ 1, 99 ~ NA)`. For more information, see details and examples. The `ifs` function checks whether one or more conditions are met and returns a value that corresponds to the first TRUE condition. `ifs` can take the place of multiple nested `ifelse` statements and is much easier to read with multiple conditions. `ifs` works in the same manner as `recode` - e. g. with formulas or with from/to notation. But conditions should be only logical and it doesn't operate on multicolumn objects.

Usage

```
if_val(x, ..., from = NULL, to = NULL)

if_val(x, from = NULL) <- value

recode(x, from = NULL) <- value

recode(x, ..., from = NULL, to = NULL)

ifs(..., from = NULL, to = NULL)

lo

hi

copy(x)

values %into% names
```

Arguments

`x`	vector/matrix/data.frame/list
`...`	sequence of formulas which describe recodings. They are used when `from`/`to` arguments are not provided.

from	list of conditions for values which should be recoded (in the same format as LHS of formulas).
to	list of values into which old values should be recoded (in the same format as RHS of formulas).
value	list with formulas which describe recodings in assignment form of function/to list if from/to notation is used.
values	object(-s) which will be assigned to names for %into% operation. %into% supports multivalue assignments. See examples.
names	name(-s) which will be given to values expression. For %into%.

Format

An object of class numeric of length 1.

Details

Input conditions - possible values for left-hand side (LHS) of formula or element of from list:

- vector/single value All values in x which equal to elements of vector in LHS will be replaced with RHS.
- function Values for which function gives TRUE will be replaced with RHS. There are some special functions for the convenience - see criteria. One of special functions is other. It means all other unrecoded values (ELSE in SPSS RECODE). All other unrecoded values will be changed to RHS of formula or appropriate element of to.
- logical vector/matrix/data.frame Values for which LHS equals to TRUE will be recoded. Logical vector will be recycled across all columns of x. If LHS is matrix/data.frame then column from this matrix/data.frame will be used for corresponding column/element of x.

Output values - possible values for right-hand side (RHS) of formula or element of to list:

- value replace elements of x. This value will be recycled across rows and columns of x.
- vector values of this vector will be replace values in corresponding position in rows of x. Vector will be recycled across columns of x.
- list/matrix/data.frame Element of list/column of matrix/data.frame will be used as a replacement value for corresponding column/element of x.
- function This function will be applied to values of x which satisfy recoding condition.There is special auxiliary function copy which just returns its argument. So in the recode it just copies old value (COPY in SPSS RECODE). See examples. copy is useful in the usual form of recode and doesn't do anything in the case of the assignment form recode() = ... because this form don't modify values which are not satisfying any of the conditions.

%into% tries to mimic SPSS 'INTO'. Values from left-hand side will be assigned to right-hand side. You can use %to% expression in the RHS of %into%. See examples. lo and hi are shortcuts for -Inf and Inf. They can be useful in expressions with %thru%, e. g. 1 %thru% hi. if_val is an alias for recode.

Value

object of same form as x with recoded values

Examples

```
# `ifs` examples
a = 1:5
b = 5:1
ifs(b>3 ~ 1)                       # c(1, 1, NA, NA, NA)
ifs(b>3 ~ 1, TRUE ~ 3)             # c(1, 1, 3, 3, 3)
ifs(b>3 ~ 1, a>4 ~ 7, TRUE ~ 3)    # c(1, 1, 3, 3, 7)
ifs(b>3 ~ a, TRUE ~ 42)            # c(1, 2, 42, 42, 42)
# some examples from SPSS manual
# RECODE V1 TO V3 (0=1) (1=0) (2, 3=-1) (9=9) (ELSE=SYSMIS)
set.seed(123)
v1  = sample(c(0:3, 9, 10), 20, replace = TRUE)
recode(v1) = c(0 ~ 1, 1 ~ 0, 2:3 ~ -1, 9 ~ 9, other ~ NA)
v1

# RECODE QVAR(1 THRU 5=1)(6 THRU 10=2)(11 THRU HI=3)(ELSE=0).
set.seed(123)
qvar = sample((-5):20, 50, replace = TRUE)
recode(qvar, 1 %thru% 5 ~ 1, 6 %thru% 10 ~ 2, 11 %thru% hi ~ 3, other ~ 0)
# the same result
recode(qvar, 1 %thru% 5 ~ 1, 6 %thru% 10 ~ 2, ge(11) ~ 3, other ~ 0)

# RECODE STRNGVAR ('A', 'B', 'C'='A')('D', 'E', 'F'='B')(ELSE=' ').
strngvar = LETTERS
recode(strngvar, c('A', 'B', 'C') ~ 'A', c('D', 'E', 'F') ~ 'B', other ~ ' ')

# RECODE AGE (MISSING=9) (18 THRU HI=1) (0 THRU 18=0) INTO VOTER.
set.seed(123)
age = sample(c(sample(5:30, 40, replace = TRUE), rep(9, 10)))
voter = recode(age, NA ~ 9, 18 %thru% hi ~ 1, 0 %thru% 18 ~ 0)
voter
# the same result with '%into%'
recode(age, NA ~ 9, 18 %thru% hi ~ 1, 0 %thru% 18 ~ 0) %into% voter2
voter2

# multiple assignment with '%into%'
#' set.seed(123)
x1 = runif(30)
x2 = runif(30)
x3 = runif(30)
# note nessesary brackets around RHS of '%into%'
recode(x1 %to% x3, gt(0.5) ~ 1, other ~ 0) %into% (x_rec_1 %to% x_rec_3)
fre(x_rec_1)
# the same operation with characters expansion
i = 1:3
recode(x1 %to% x3, gt(0.5) ~ 1, other ~ 0) %into% subst('x_rec2_`i`')
fre(x_rec2_1)

# example with function in RHS
set.seed(123)
a = rnorm(20)
# if a<(-0.5) we change it to absolute value of a (abs function)
```

```
recode(a, lt(-0.5) ~ abs, other ~ copy)

# the same example with logical criteria
recode(a, a<(-.5) ~ abs, other ~ copy)

# replace with specific value for each column
# we replace values greater than 0.75 with column max and values less than 0.25 with column min
# and NA with column means
# make data.frame
set.seed(123)
x1 = runif(30)
x2 = runif(30)
x3 = runif(30)
x1[sample(30, 10)] = NA # place 10 NA's
x2[sample(30, 10)] = NA # place 10 NA's
x3[sample(30, 10)] = NA # place 10 NA's
dfs = data.frame(x1, x2, x3)

#replacement. Note the necessary transpose operation
recode(dfs,
        lt(0.25) ~ t(min_col(dfs)),
        gt(0.75) ~ t(max_col(dfs)),
        NA ~ t(mean_col(dfs)),
        other ~ copy
      )

# replace NA with row means
# some rows which contain only NaN's remain unchanged because mean_row for them also is NaN
recode(dfs, NA ~ mean_row(dfs), other ~ copy)

# some of the above examples with from/to notation

set.seed(123)
v1  = sample(c(0:3,9,10), 20, replace = TRUE)
# RECODE V1 TO V3 (0=1) (1=0) (2,3=-1) (9=9) (ELSE=SYSMIS)
fr = list(0, 1, 2:3, 9, other)
to = list(1, 0, -1, 9, NA)
recode(v1, from = fr) = to
v1

# RECODE QVAR(1 THRU 5=1)(6 THRU 10=2)(11 THRU HI=3)(ELSE=0).
fr = list(1 %thru% 5, 6 %thru% 10, ge(11), other)
to = list(1, 2, 3, 0)
recode(qvar, from = fr, to = to)

# RECODE STRNGVAR ('A','B','C'='A')('D','E','F'='B')(ELSE=' ').
fr = list(c('A','B','C'), c('D','E','F') , other)
to = list("A", "B", " ")
recode(strngvar, from = fr, to = to)

# RECODE AGE (MISSING=9) (18 THRU HI=1) (0 THRU 18=0) INTO VOTER.
fr = list(NA, 18 %thru% hi, 0 %thru% 18)
to = list(9, 1, 0)
```

```
voter = recode(age, from = fr, to = to)
voter
```

`info` *Provides variables description for dataset*

Description

`info` returns data.frame with variables description and some summary statistics. Resulting data.frame mainly intended to keep in front of eyes in RStudio viewer or to be saved as csv to view in the spreadsheet software as reference about working dataset.

Usage

```
info(x, stats = TRUE, frequencies = TRUE, max_levels = 10)
```

Arguments

`x`	vector/factor/list/data.frame.
`stats`	Logical. Should we calculate summary for each variable?
`frequencies`	Logical. Should we calculate frequencies for each variable? This calculation can take significant amount of time for large datasets.
`max_levels`	Numeric. Maximum levels for using in frequency calculations. Levels above this value will convert to 'Other values'.

Value

data.frame with following columns: Name, Class, Length, NotNA, NA, Distincts, Label, ValueLabels, Min., 1st Qu., Median, Mean, 3rd Qu., Max., Frequency.

Examples

```
data(mtcars)
var_lab(mtcars$am) = "Transmission"
val_lab(mtcars$am) = c("Automatic"=0, "Manual"=1)
info(mtcars, max_levels = 5)
```

keep *Keep or drop elements by name/criteria in data.frame/matrix*

Description

keep selects variables/elements from data.frame by their names or by criteria (see criteria). except drops variables/elements from data.frame by their names or by criteria. Names at the top-level can be unquoted (non-standard evaluation). For standard evaluation of parameters you can surround them by round brackets. See examples. Methods for list will apply keep/except to each element of the list separately. .keep/.except are versions which works with default_dataset.

Usage

```
keep(data, ...)

.keep(...)

except(data, ...)

.except(...)
```

Arguments

data	data.frame/matrix/list
...	column names of type character/numeric or criteria/logical functions

Value

object of the same type as data

Examples

```
data(iris)
keep(iris, Sepal.Length, Sepal.Width)
except(iris, Species)

keep(iris, Species, other()) # move 'Species' to the first position
keep(iris, to("Petal.Width")) # keep all columns up to 'Species'

except(iris, perl("^Petal")) # remove columns which names start with 'Petal'

except(iris, 5) # remove fifth column

data(mtcars)
keep(mtcars, from("mpg") & to("qsec")) # keep columns from 'mpg' to 'qsec'
keep(mtcars, mpg %to% qsec) # the same result

 # standard and non-standard evaluation
```

```
 many_vars = c("am", "vs", "cyl")
 ## Not run:
 keep(mtcars, many_vars) # error - names not found: 'many_vars'

## End(Not run)
 keep(mtcars, (many_vars)) # ok

# character expansion
dfs = data.frame(
     a = 10 %r% 5,
     b_1 = 11 %r% 5,
     b_2 = 12 %r% 5,
     b_3 = 12 %r% 5,
     b_4 = 14 %r% 5,
     b_5 = 15 %r% 5
 )
 i = 1:5
 keep(dfs, b_1 %to% b_5)
 keep(dfs, subst("b_`i`")) # the same result
```

`match_row` *Match finds value in rows or columns/index returns value by index from rows or columns*

Description

`match` finds value in rows or columns. `index` returns value by index from row or column. One can use functions as criteria for `match`. In this case position of first value on which function equals to TRUE will be returned. For convenience there are special predefined functions - see criteria. If value is not found then NA will be returned.

Usage

```
match_row(criterion, ...)

match_col(criterion, ...)

index_row(index, ...)

index_col(index, ...)
```

Arguments

`criterion`	Vector of values to be matched, or function.
`...`	data. Vectors, matrixes, data.frames, lists. Shorter arguments will be recycled.
`index`	vector of positions in rows/columns from which values should be returned.

Value

vector with length equals to number of rows for *_row and equals to number of columns for *_col.

Examples

```
# toy data
v1 = 1:3
v2 = 2:4
v3 = 7:5

# postions of 1,3,5 in rows
match_row(c(1, 3, 5), v1, v2, v3) # 1:3
# postions of 1,3,5 in columns
match_col(1, v1, v2, v3) # c(v1 = 1, v2 = NA, v3 = NA)

# postion of first value greater than 2
ix = match_row(gt(2), v1, v2, v3)
ix # c(3,2,1)
# return values by result of previous 'match_row'
index_row(ix, v1, v2, v3) # c(7,3,3)

# the same actions with data.frame
dfs = data.frame(v1, v2, v3)

# postions of 1,3,5 in rows
match_row(c(1, 3, 5), dfs) # 1:3
# postions of 1,3,5 in columns
match_col(1, dfs) # c(v1 = 1, v2 = NA, v3 = NA)

# postion of first value greater than 2
ix = match_row(gt(2), dfs)
ix # c(3,2,1)
# return values by result of previous 'match_row'
index_row(ix, dfs) # c(7,3,3)
```

`merge.etable` *Merge two tables/data.frames*

Description

`%merge%` is infix shortcut for base merge with `all.x = TRUE` and `all.y = FALSE` (left join). There is also special method for combining results of `cro_*` and `fre`. For them `all = TRUE` (full join). It allows make complex tables from simple ones. See examples. Strange result is possible if one or two arguments have duplicates in first column (column with labels).

Usage

```
x %merge% y
```

Arguments

x	data.frame or results of fre/cro_*/table_*
y	data.frame or results of fre/cro_*/table_*

Value

data.frame

See Also

fre, cro, cro, cro_fun, merge

Examples

```
data(mtcars)
# apply labels
mtcars = apply_labels(mtcars,
                mpg = "Miles/(US) gallon",
                cyl = "Number of cylinders",
                disp = "Displacement (cu.in.)",
                hp = "Gross horsepower",
                drat = "Rear axle ratio",
                wt = "Weight (lb/1000)",
                qsec = "1/4 mile time",
                vs = "V/S",
                vs = c("V-engine" = 0, "Straight engine" = 1),
                am = "Transmission (0 = automatic, 1 = manual)",
                am = c(automatic = 0, manual = 1),
                gear = "Number of forward gears",
                carb = "Number of carburetors"
)

# table by 'am'
tab1 = calculate(mtcars, cro_cpct(gear, am))
# table with percents
tab2 = calculate(mtcars, cro_cpct(gear, vs))

# combine tables
tab1 %merge% tab2

# complex tables
# table with counts
counts = calculate(mtcars, cro(list(vs, am, gear, carb), list("Count")))
# table with percents
percents = calculate(mtcars, cro_cpct(list(vs, am, gear, carb), list("Column, %")))

# combine tables
counts %merge% percents
```

modify *Modify data.frame/modify subset of the data.frame*

Description

- compute evaluates expression expr in the context of data.frame data and return original data possibly modified. It works similar to within in base R but try to return new variables in order of their occurrence in the expression and make available full-featured %to% and .N in the expressions. See vars.
- calculate evaluates expression expr in the context of data.frame data and return value of the evaluated expression. It works similar to with in base R but make available full-featured %to% and .N in the expressions. See vars. Function use_labels is shortcut for calculate with argument use_labels set to TRUE.
- do_if modifies only rows for which cond equals to TRUE. Other rows remain unchanged. Newly created variables also will have values only in rows for which cond have TRUE. There will be NA's in other rows. This function tries to mimic SPSS "DO IF(). ... END IF." statement.

There is a special constant .N which equals to number of cases in data for usage in expression inside compute/calculate. Inside do_if .N gives number of rows which will be affected by expressions. For parametrization (variable substitution) see .. or examples. Sometimes it is useful to create new empty variable inside compute. You can use .new_var function for this task. This function creates variable of length .N filled with NA. See examples. modify is an alias for compute, modify_if is an alias for do_if and calc is an alias for calculate.

Usage

```
modify(data, expr)

compute(data, expr)

modify_if(data, cond, expr)

do_if(data, cond, expr)

calculate(data, expr, use_labels = FALSE)

use_labels(data, expr)

calc(data, expr, use_labels = FALSE)

data %calc% expr

data %use_labels% expr

data %calculate% expr
```

Arguments

data	data.frame/list of data.frames. If data is list of data.frames then expression expr will be evaluated inside each data.frame separately.
expr	expression that should be evaluated in the context of data.frame data
cond	logical vector or expression. Expression will be evaluated in the context of the data.
use_labels	logical. Experimental feature. If it equals to TRUE then we will try to replace variable names with labels. So many base R functions which show variable names will show labels.

Value

compute and do_if functions return modified data.frame/list of modified data.frames, calculate returns value of the evaluated expression/list of values.

Examples

```
dfs = data.frame(
    test = 1:5,
    a = rep(10, 5),
    b_1 = rep(11, 5),
    b_2 = rep(12, 5),
    b_3 = rep(13, 5),
    b_4 = rep(14, 5),
    b_5 = rep(15, 5)
)


# compute sum of b* variables and attach it to 'dfs'
compute(dfs, {
    b_total = sum_row(b_1 %to% b_5)
    var_lab(b_total) = "Sum of b"
    random_numbers = runif(.N) # .N usage
})

# calculate sum of b* variables and return it
calculate(dfs, sum_row(b_1 %to% b_5))


# set values to existing/new variables
compute(dfs, {
    (b_1 %to% b_5) %into% subst('new_b`1:5`')
})

# .new_var usage
compute(dfs, {
    new_var = .new_var()
    new_var[1] = 1 # this is not possible without preliminary variable creation
})
```

```
# conditional modification
do_if(dfs, test %in% 2:4, {
    a = a + 1
    b_total = sum_row(b_1 %to% b_5)
    random_numbers = runif(.N) # .N usage
})


# variable substitution
name1 = "a"
name2 = "new_var"

# example with short notation but it can be applied only for simple cases -
# when 'name' is vector of length 1
compute(dfs, {
     ..$name2 = ..$name1*2
})

compute(dfs, {
     for(name1 in paste0("b_", 1:5)){
         name2 = paste0("new_", name1)
         ..$name2 = ..$name1*2
     }
     rm(name1, name2) # we don't need this variables as columns in 'dfs'
})

# square brackets notation
compute(dfs, {
     ..[(name2)] = ..[(name1)]*2
})

compute(dfs, {
     for(name1 in paste0("b_", 1:5)){
         ..[paste0("new_", name1)] = ..$name1*2
     }
     rm(name1) # we don't need this variable as column in 'dfs'
})

# '..$' doesn't work for case below so we need to use square brackets form
name1 = paste0("b_", 1:5)
name2 = paste0("new_", name1)
compute(dfs, {
     for(i in 1:5){
         ..[name2[i]] = ..[name1[i]]*3
     }
     rm(i) # we don't need this variable as column in 'dfs'
})

# 'use_labels' examples. Utilization of labels in base R.
data(mtcars)
mtcars = apply_labels(mtcars,
                      mpg = "Miles/(US) gallon",
                      cyl = "Number of cylinders",
```

```
                        disp = "Displacement (cu.in.)",
                        hp = "Gross horsepower",
                        drat = "Rear axle ratio",
                        wt = "Weight (lb/1000)",
                        qsec = "1/4 mile time",
                        vs = "Engine",
                        vs = c("V-engine" = 0,
                               "Straight engine" = 1),
                        am = "Transmission",
                        am = c("Automatic" = 0,
                               "Manual"=1),
                        gear = "Number of forward gears",
                        carb = "Number of carburetors"
)

use_labels(mtcars, table(am, vs))

## Not run:
use_labels(mtcars, plot(mpg, hp))

## End(Not run)

mtcars %>%
       use_labels(lm(mpg ~ disp + hp + wt)) %>%
       summary()
```

`mrset` *Create multiple response set/multiple dichotomy set from variables*

Description

These functions are intended for usage with tables - tables, cro, cro_fun. Result of `mrset` is considered as muliple-response set with category encoding and result of `mdset` is considered as multiple response set with dichotomy (dummy) encoding e. g. with 0 or 1 in the each column. Each column in the `dichotomy` is indicator of absense or presense of particular feature. Both functions don't convert its arguments to anything - it is supposed that arguments already have appropriate encoding. For conversation see as.dichotomy or as.category.

Usage

```
mrset(..., label = NULL)

mdset(..., label = NULL)
```

Arguments

`...`	variables
`label`	character optional label for multiple response set

Value

data.frame of class `category/dichotomy`

See Also

as.dichotomy, as.category

Examples

```
data(product_test)
a1 = calculate(product_test, mrset(a1_1 %to% a1_6))
```

`names2labels` *Replace data.frame/list names with corresponding variables labels.*

Description

`names2labels` replaces data.frame/list names with corresponding variables labels. If there are no labels for some variables their names remain unchanged. `n2l` is just shortcut for `names2labels`.

Usage

```
names2labels(x, exclude = NULL, keep_names = FALSE)

n2l(x, exclude = NULL, keep_names = FALSE)
```

Arguments

`x`	data.frame/list.
`exclude`	logical/integer/character columns which names should be left unchanged. Only applicable to list/data.frame.
`keep_names`	logical. If TRUE original column names will be kept with labels. Only applicable to list/data.frame.

Value

Object of the same type as x but with variable labels instead of names.

See Also

values2labels, val_lab, var_lab

Examples

```
data(mtcars)
mtcars = modify(mtcars,{
                var_lab(mpg) = "Miles/(US) gallon"
                var_lab(cyl) = "Number of cylinders"
                var_lab(disp) = "Displacement (cu.in.)"
                var_lab(hp) = "Gross horsepower"
                var_lab(drat) = "Rear axle ratio"
                var_lab(wt) = "Weight (lb/1000)"
                var_lab(qsec) = "1/4 mile time"
                var_lab(vs) = "V/S"
                var_lab(am) = "Transmission (0 = automatic, 1 = manual)"
                var_lab(gear) = "Number of forward gears"
                var_lab(carb) = "Number of carburetors"
})

# without original names
# note: we exclude dependent variable 'mpg' from conversion to use its short name in formula
summary(lm(mpg ~ ., data = names2labels(mtcars, exclude = "mpg")))
# with names
summary(lm(mpg ~ ., data = names2labels(mtcars, exclude = "mpg", keep_names = TRUE)))
```

---

na_if *Replace certain values with NA*

Description

There are following options for `value`:

- `vector` Vector of values which should be replaced with `NA` in `x`.
- `logical vector/matrix/data.frame` NA's will be set in places where `value` is TRUE. `value` will be recycled if needed.
- `function` NA's will be set in places where `value(x)` is TRUE. Function will be applied columnwise. Additionally, there are special functions for common cases of comparison. For example `na_if(my_var, gt(98))` will replace all values which are greater than 98 in `my_var` with NA. For detailed description of special functions see criteria

`mis_val` is an alias for the `na_if` with absolutely the same functionality.

Usage

```
na_if(x, value)

na_if(x) <- value

x %na_if% value

mis_val(x, value)
```

```
mis_val(x) <- value

x %mis_val% value
```

Arguments

x	vector/matrix/data.frame/list
value	vector/matrix/data.frame/function

Value

x with NA's instead of value

See Also

For reverse operation see if_na, if_val for more general recodings.

Examples

```
a = c(1:5, 99)

# 99 to NA
na_if(a, 99)    # c(1:5, NA)

a %na_if% 99    # same result

# values which greater than 5 to NA
na_if(a, gt(5)) # c(1:5, NA)

set.seed(123)
dfs = data.frame(
      a = c("bad value", "bad value", "good value", "good value", "good value"),
      b = runif(5)
)

# rows with 'bad value' will be filled with NA
# logical argument and recycling by columns
na_if(dfs, dfs$a=="bad value")

a = rnorm(50)
# values greater than 1 or less than -1 will be set to NA
# special functions usage
na_if(a, lt(-1) | gt(1))

# values inside [-1, 1] to NA
na_if(a, -1 %thru% 1)
```

`nest` *Compute nested variable(-s) from several variables*

Description

`nest` mainly intended for usage with table functions such as cro. See examples. `%nest%` is infix version of this function. You can apply `nest` on multiple-response variables/list of variables and data.frames.

Usage

```
nest(...)

x %nest% y
```

Arguments

`...`	vectors/data.frames/lists
`x`	vector/data.frame/list
`y`	vector/data.frame/list

Value

vector/data.frame/list

See Also

See also interaction

Examples

```
data(mtcars)

mtcars = apply_labels(mtcars,
                      cyl = "Number of cylinders",
                      vs = "Engine",
                      vs = num_lab("
                             0 V-engine
                             1 Straight engine
                             "),
                      am = "Transmission",
                      am = num_lab("
                             0 Automatic
                             1 Manual
                             "),
                      carb = "Number of carburetors"
)
```

```
with(mtcars, cro(cyl, am %nest% vs))

# three variables
with(mtcars, cro(am %nest% vs %nest% carb, cyl))

# the same with usual version
with(mtcars, cro(cyl, nest(am, vs)))

# three variables
with(mtcars, cro(nest(am, vs, carb), cyl))
```

`prepend_values` *Prepend values/variable names to value/variable labels*

Description

These functions add values/variable names as prefixes to value/variable labels. Functions which start with `tab_` intended for usage inside table creation sequences. See examples and tables. It is recommended to use `tab_prepend_*` at the start of sequence of tables creation. If you use it in the middle of the sequence then previous statements will not be affected.

Usage

```
prepend_values(x)

prepend_names(x)

prepend_all(x)

tab_prepend_values(data)

tab_prepend_names(data)

tab_prepend_all(data)
```

Arguments

`x`	vector/data.frame. `prepend_names` can be applied only to data.frames.
`data`	data.frame/intermediate result of tables construction. See tables.

Value

original object with prepended names/values to labels

Examples

```
data(mtcars)
mtcars = apply_labels(mtcars,
                      mpg = "Miles/(US) gallon",
                      cyl = "Number of cylinders",
                      disp = "Displacement (cu.in.)",
                      hp = "Gross horsepower",
                      drat = "Rear axle ratio",
                      wt = "Weight (lb/1000)",
                      qsec = "1/4 mile time",
                      vs = "Engine",
                      vs = c("V-engine" = 0,
                             "Straight engine" = 1),
                      am = "Transmission",
                      am = c("Automatic" = 0,
                             "Manual"=1),
                      gear = "Number of forward gears",
                      carb = "Number of carburetors"
)

# prepend names and 'cro_cpct'
mtcars %>%
       prepend_names %>%
       calculate(
              cro_cpct(list(cyl, gear), list(total(), vs, am))
          )

# prepend values to value labels
mtcars %>%
   tab_prepend_values %>%
   tab_cols(total(), vs, am) %>%
   tab_cells(cyl, gear) %>%
   tab_stat_cpct() %>%
   tab_pivot()

# prepend names and labels
mtcars %>%
   tab_prepend_all %>%
   tab_cols(total(), vs, am) %>%
   tab_cells(cyl, gear) %>%
   tab_stat_cpct() %>%
   tab_pivot()

# variable in rows without prefixes
mtcars %>%
   tab_cells(cyl, gear) %>%
   tab_prepend_all %>%
   tab_cols(total(), vs, am) %>%
   tab_stat_cpct() %>%
   tab_pivot()
```

product_test *Data from product test of chocolate confectionary*

Description

It is truncated dataset with data from product test of two samples of chocolate sweets. 150 respondents tested two kinds of sweets (codenames: VSX123 and SDF546). Sample was divided into two groups (cells) of 75 respondents in each group. In cell 1 product VSX123 was presented first and then SDF546. In cell 2 sweets were presented in reversed order. Questions about respondent impressions about first product are in the block A (and about second tested product in the block B). At the end of the questionnaire there is a question about preferences between sweets.

Usage

```
product_test
```

Format

A data frame with 150 rows and 18 variables:

id Respondent Id.

cell First tested product (cell number).

s2a Age.

a1_1 What did you like in these sweets? Multiple response. First tested product.

a1_2 (continue) What did you like in these sweets? Multiple response. First tested product.

a1_3 (continue) What did you like in these sweets? Multiple response. First tested product.

a1_4 (continue) What did you like in these sweets? Multiple response. First tested product.

a1_5 (continue) What did you like in these sweets? Multiple response. First tested product.

a1_6 (continue) What did you like in these sweets? Multiple response. First tested product.

a22 Overall liking. First tested product.

b1_1 What did you like in these sweets? Multiple response. Second tested product.

b1_2 (continue) What did you like in these sweets? Multiple response. Second tested product.

b1_3 (continue) What did you like in these sweets? Multiple response. Second tested product.

b1_4 (continue) What did you like in these sweets? Multiple response. Second tested product.

b1_5 (continue) What did you like in these sweets? Multiple response. Second tested product.

b1_6 (continue) What did you like in these sweets? Multiple response. Second tested product.

b22 Overall liking. Second tested product.

c1 Preferences.

prop *Compute proportions from numeric vector/matrix/data.frame*

Description

prop returns proportion to sum of entire x. prop_col returns proportion to sum of each column of x. prop_row returns proportion to sum of each row of x. Non-numeric columns in the data.frame are ignored. NA's are also ignored.

Usage

```
prop(x)

prop_col(x)

prop_row(x)
```

Arguments

x	numeric vector/matrix/data.frame

Value

the same structure as x but with proportions of original values from sum of original values.

Examples

```
a = c(25, 25, NA)
prop(a)

# data.frame with non-numeric columns
fac = factor(c("a", "b", "c"))
char = c("a", "b", "c")
dat = as.POSIXct("2016-09-27")
a = dtfrm(fac, a = c(25, 25, NA), b = c(100, NA, 50), char, dat)

prop(a)
prop_row(a)
prop_col(a)

# the same as result as with 'prop.table'
tbl = table(state.division, state.region)

prop(tbl)
prop_row(tbl)
prop_col(tbl)
```

qc *Create vector of characters from unquoted strings (variable names)*

Description

In many cases one need to address variables in list/data.frame in such manner: `dfs[ , c("var1", "var2", "var3")]`. `qc` ("quoted c") is a shortcut for the such cases to reduce keystrokes. With `qc` you can write: `dfs[ , qc(var1, var2, var3)]`. `subst` is simple string interpolation function. It searches in its arguments expressions in backticks (`), evaluate it and substitute it with result of evaluation. See examples.

Usage

```
qc(...)

subst(...)
```

Arguments

`...`	characters in `subst`/unquoted names of variables in `qc`

Value

Vector of characters

Examples

```
## qc
qc(a, b, c)
identical(qc(a, b, c), c("a", "b", "c"))

mtcars[, qc(am, mpg, gear)]

## subst
i = 1:5
subst("q`i`")

i = 1:3
j = 1:3
subst("q1_`i`_`j`")

data(iris)
subst("'iris' has `nrow(iris)` rows.")
```

read_spss *Read an SPSS Data File*

Description

read_spss reads data from a file stored in SPSS *.sav format. It returns data.frame and never converts string variables to factors. Also it prepares SPSS values/variables labels for working with val_lab/var_lab functions. User-missings values are ignored. read_spss is simple wrapper around read.spss function from package foreign.

Usage

```
read_spss(file, reencode = TRUE)

read_spss_to_list(file, reencode = TRUE)
```

Arguments

file	Character string: the name of the file or URL to read.
reencode	logical: should character strings be re-encoded to the current locale. The default is TRUE. NA means to do so in a UTF-8 locale, only. Alternatively, a character string specifying an encoding to assume for the file.

Value

read_spss returns data.frame.

read_spss_to_list returns list of variables from SPSS files.

See Also

read.spss in package foreign, val_lab, var_lab

Examples

```
## Not run:

w = read_spss("project_123.sav") # to data.frame
list_w = read_spss_to_list("project_123.sav") # to list


## End(Not run)
```

ref *Auxiliary functions to pass arguments to function by reference*

Description

These two functions aimed to simplify build functions with side-effects (e. g. for modifying variables in place). Of cause it is not the R way of doing things but sometimes it can save several keystrokes.

Usage

```
ref(x)

ref(x) <- value
```

Arguments

x	Reference to variable, it is formula, ~var_name.
value	Value that should be assigned to modified variable.

Details

To create reference to variable one can use formula: `b = ~a`. `b` is reference to `a`. So `ref(b)` returns value of `a` and `ref(b) = new_val` will modify `a`. If argument `x` of these functions is not formula then these functions have no effect e. g. `ref(a)` is identical to `a` and after `ref(a) = value` `a` is identical to `value`. It is not possible to use function as argument `x` in assignment form. For example, `ref(some_function(x)) = some_value` will rise error. Use `y = some_function(x); ref(y) = some_value` instead.

Value

`ref` returns value of referenced variable. `ref<-` modifies referenced variable.

Examples

```
# Simple example
a = 1:3
b = ~a  # b is reference to 'a'
identical(ref(b),a) # TRUE

ref(b)[2] = 4 # here we modify 'a'
identical(a, c(1,4,3)) # TRUE

# usage inside function

# top 10 rows
head10 = function(x){
 ds = head(ref(x), 10)
```

```
 ref(x) = ds
 invisible(ds) # for usage without references
}

data(iris)
ref_to_iris = ~iris
head10(ref_to_iris) # side-effect
nrow(iris) # 10

# argument is not formula - no side-effect
data(mtcars)
mtcars10 = head10(mtcars)

nrow(mtcars10) # 10
nrow(mtcars) # 32
```

`sheet` *Make data.frame without conversion to factors and without fixing names*

Description

`sheet` and `as.sheet` are shortcuts to `data.frame` and `as.data.frame` with stringsAsFactors = FALSE, check.names = FALSE. `lst` creates list with names. `.sheet`, `.lst` the same as above but work in the scope of default dataset.

Usage

```
sheet(...)

as.sheet(x, ...)

dtfrm(...)

as.dtfrm(x, ...)

lst(...)

.dtfrm(...)

.sheet(...)

.lst(...)
```

Arguments

`...`	objects, possibly named
`x`	object to be coerced to data.frame

Value

data.frame/list

See Also

default_dataset, data.frame, as.data.frame, list

Examples

```
# see the difference
df1 = data.frame(a = letters[1:3], "This is my long name" = 1:3)
df2 = sheet(a = letters[1:3], "This is my long name" = 1:3)

str(df1)
str(df2)

# lst
a = 1:3
b = 3:1

list1 = list(a, b)
list2 = lst(a, b)

str(list1)
str(list2)

data(iris)
default_dataset(iris)

.sheet(Sepal.Width,  Sepal.Length)
.lst(Sepal.Width,  Sepal.Length)
```

---

sort_asc *Sort data.frames/matrices/vectors*

Description

`sort_asc` sorts in ascending order and `sort_desc` sorts in descending order. `.sort_asc`/`.sort_desc` are versions for working with default_dataset.

Usage

```
sort_asc(data, ..., na.last = FALSE)

.sort_asc(..., na.last = FALSE)

sort_desc(data, ..., na.last = TRUE)
```

```
.sort_desc(..., na.last = TRUE)
```

Arguments

data	data.frame/matrix/vector
...	character/numeric or criteria/logical functions (see criteria). Column names/numbers for data.frame/matrix by which object will be sorted. Names at the top-level can be unquoted (non-standard evaluation). For standard evaluation of parameters you can surround them by round brackets. See examples. Ignored for vectors.
na.last	for controlling the treatment of NAs. If TRUE, missing values in the data are put last; if FALSE, they are put first; if NA, they are removed.

Value

sorted data

Examples

```
data(mtcars)
sort_asc(mtcars, mpg)
sort_asc(mtcars, cyl, mpg) # by two column

# same results with column nums
sort_asc(mtcars, 1)
sort_asc(mtcars, 2:1) # by two column
sort_asc(mtcars, 2, 1) # by two column

# call with parameter
sorting_columns = c("cyl", "mpg")
sort_asc(mtcars, (sorting_columns))
```

---

split_labels *Split character vector to matrix/split columns in data.frame*

Description

split_labels/split_columns are auxiliary functions for post-processing tables resulted from cro/cro_fun and etc. In these tables all labels collapsed in the first column with "|" separator. split_columns split first column into multiple columns with separator (split argument). split_table_to_df split first column of table and column names. Result of this operation is data.frame with character columns.

Usage

```
split_labels(x, remove_repeated = TRUE, split = "\\|", fixed = FALSE,
  perl = FALSE)

split_columns(data, columns = 1, remove_repeated = TRUE, split = "\\|",
  fixed = FALSE, perl = FALSE)

split_table_to_df(data, digits = get_expss_digits(), remove_repeated = TRUE,
  split = "\\|", fixed = FALSE, perl = FALSE)

make_subheadings(data, number_of_columns = 1)
```

Arguments

x	character vector which will be split
remove_repeated	logical. Default is TRUE. Should we remove repeated labels?
split	character vector (or object which can be coerced to such) containing regular expression(s) (unless fixed = TRUE) to use for splitting.
fixed	logical. If TRUE match split exactly, otherwise use regular expressions. Has priority over perl.
perl	logical. Should Perl-compatible regexps be used?
data	data.frame vector which will be split
columns	character/numeric/logical columns in the data.frame data which should be split
digits	numeric. How many digits after decimal point should be left in split_table_to_df?
number_of_columns	integer. Number of columns from row labels which will be used as subheadings in table.

Value

split_labels returns character matrix, split_columns returns data.frame with columns replaced by possibly multiple columns with split labels. split_table_to_df returns data.frame with character columns.

See Also

strsplit

Examples

```
data(mtcars)

# apply labels
mtcars = apply_labels(mtcars,
    cyl = "Number of cylinders",
    vs = "Engine",
```

```
    vs = c("V-engine" = 0,
                   "Straight engine" = 1),
    am = "Transmission",
    am = c(automatic = 0,
                   manual=1),
    gear = "Number of forward gears",
    carb = "Number of carburetors"
)

# all row labels in the first column
tabl = mtcars %>%
       calculate(cro_cpct(list(cyl, gear, carb), list(total(), vs, am)))

tabl # without subheadings

make_subheadings(tabl) # with subheadings

split_labels(tabl[[1]])
split_labels(colnames(tabl))

# replace first column with new columns
split_columns(tabl) # remove repeated

split_columns(tabl, remove_repeated = FALSE)

split_columns(tabl)

split_table_to_df(tabl)

split_table_to_df(tabl)
```

`split_separate` *Splits data.frame into list of data.frames that can be analyzed separately*

Description

Splits data.frame into list of data.frames that can be analyzed separately. These data.frames are sets of cases that have the same values for the specified split variables. Any missing values in split variables are dropped together with the corresponding values of `data`. `split_off` works with lists of data.frames or objects that can be coerced to data.frame and assumed to have compatible structure. Resulting rows will be sorted in order of the split variables.

Usage

```
split_separate(data, ..., drop = TRUE)

split_off(data, groups = NULL, rownames = NULL)
```

Arguments

data	data.frame for split_separate/list for split_off
...	unquoted variables names (see keep) by which data will be split into list.
drop	should we drop combination of levels with zero observation? TRUE by default.
groups	character If it is not NULL then we add list names as variable to result of split_off with the name specified by groups. If it is TRUE then name will be .groups.
rownames	character If it is not NULL then we add data.frames rownames as variable to result of split_off with the name specified by rownames. If it is TRUE then name will be .rownames.

Value

split_separate returns list of data.frames/split_off returns data.frame

See Also

split, compute, calculate, do_repeat, where

Examples

```
# example from base R 'split'
data(airquality)
airquality2 = airquality %>%
    split_separate(Month) %>%
    compute({
        Ozone_zscore = scale(Ozone)
    }) %>%
    split_off()

head(airquality2)

# usage of 'groups', 'rownames'
data(mtcars)
# add labels to dataset
mtcars %>%
    apply_labels(mpg = "Miles/(US) gallon",
                 disp = "Displacement (cu.in.)",
                 wt = "Weight",
                 hp = "Gross horsepower",
                 vs = "Engine",
                 vs = num_lab("
                                  0 V-engine
                                  1 Straight engine
                                  "),

                 am = "Transmission",
                 am = num_lab("
                                  0 Automatic
                                  1 Manual
                                  ")
```

```
) %>%
split_separate(am, vs) %>%
use_labels({
    res = lm(mpg ~ hp + disp + wt)
    cbind(Coef. = coef(res), confint(res))
}) %>%
split_off(groups = TRUE, rownames = "variable")
```

---

sum_row *Compute sum/mean/sd/median/max/min/custom function on rows/columns*

Description

These functions are intended for usage inside modify, modify_if, with and within functions. sum/mean/sd/median/max/min always omits NA. `any_in_*` checks existence of any TRUE in each row/column. It is equivalent of any applied to each row/column. `all_in_*` is equivalent of all applied to each row/column. They don't remove NA.

Usage

```
sum_row(...)

sum_col(...)

mean_row(...)

mean_col(...)

sd_row(...)

sd_col(...)

median_row(...)

median_col(...)

max_row(...)

max_col(...)

min_row(...)

min_col(...)

apply_row(fun, ...)

apply_col(fun, ...)
```

```
any_in_row(...)

any_in_col(...)

all_in_row(...)

all_in_col(...)
```

Arguments

...	data. Vectors, matrixes, data.frames, list. Shorter arguments will be recycled.
fun	custom function that will be applied to ...

Value

All functions except `apply_*` return numeric vector of length equals the number of argument columns/rows. Value of `apply_*` depends on supplied `fun` function.

See Also

modify, modify_if, %to%, count_if, sum_if, mean_if, median_if, sd_if, min_if, max_if

Examples

```
## Inside example
iris = modify(iris,{
  new_median = median_row(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)
  new_mean = mean_row(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)
  })

dfs = data.frame(
    test = 1:5,
    aa = rep(10, 5),
    b_ = rep(20, 5),
    b_1 = rep(11, 5),
    b_2 = rep(12, 5),
    b_4 = rep(14, 5),
    b_5 = rep(15, 5)
)

# calculate sum of b* variables
modify(dfs, {
    b_total = sum_row(b_, b_1 %to% b_5)
})

# conditional modification
modify_if(dfs, test %in% 2:4, {
    b_total = sum_row(b_, b_1 %to% b_5)
})
```

```
# Examples from rowSums/colSums manual.
## Compute row and column sums for a matrix:
x = cbind(x1 = 3, x2 = c(4:1, 2:5))
sum_row(x); sum_col(x)
dimnames(x)[[1]] <- letters[1:8]
sum_row(x); sum_col(x); mean_row(x); mean_col(x)
```

`tables` *Functions for custom tables construction*

Description

Table construction consists of at least of three functions chained with `magrittr` pipe operator: %>%. At first we need to specify variables for which statistics will be computed with `tab_cells`. Secondary, we calculate statistics with one of `tab_stat_*` functions. And last, we finalize table creation with `tab_pivot`: `dataset %>% tab_cells(variable) %>% tab_stat_cases() %>% tab_pivot()`. After that we can optionally sort table with tab_sort_asc, drop empty rows/columns with drop_rc and transpose with `tab_transpose`. Generally, table is just a data.frame so we can use arbitrary operations on it. Statistic is always calculated with the last cell, column/row variables, weight, missing values and subgroup. To define new cell/column/row variables we can call appropriate function one more time. `tab_pivot` defines how we combine different statistics and where statistic labels will appear - inside/outside rows/columns. See examples. For significance testing see significance.

Usage

```
tab_cols(data, ...)

tab_cells(data, ...)

tab_rows(data, ...)

tab_weight(data, weight = NULL)

tab_mis_val(data, ...)

tab_total_label(data, ...)

tab_total_statistic(data, ...)

tab_total_row_position(data, total_row_position = c("below", "above", "none"))

tab_subgroup(data, subgroup = NULL)

tab_row_label(data, ..., label = NULL)

tab_stat_fun(data, ..., label = NULL, unsafe = FALSE)
```

```
tab_stat_mean_sd_n(data, weighted_valid_n = FALSE, labels = c("Mean",
  "Std. dev.", ifelse(weighted_valid_n, "Valid N", "Unw. valid N")),
  label = NULL)

tab_stat_mean(data, label = "Mean")

tab_stat_median(data, label = "Median")

tab_stat_se(data, label = "S. E.")

tab_stat_sum(data, label = "Sum")

tab_stat_min(data, label = "Min.")

tab_stat_max(data, label = "Max.")

tab_stat_sd(data, label = "Std. dev.")

tab_stat_valid_n(data, label = "Valid N")

tab_stat_unweighted_valid_n(data, label = "Unw. valid N")

tab_stat_fun_df(data, ..., label = NULL, unsafe = FALSE)

tab_stat_cases(data, total_label = NULL, total_statistic = "u_cases",
  total_row_position = c("below", "above", "none"), label = NULL)

tab_stat_cpct(data, total_label = NULL, total_statistic = "u_cases",
  total_row_position = c("below", "above", "none"), label = NULL)

tab_stat_cpct_responses(data, total_label = NULL,
  total_statistic = "u_responses", total_row_position = c("below", "above",
  "none"), label = NULL)

tab_stat_tpct(data, total_label = NULL, total_statistic = "u_cases",
  total_row_position = c("below", "above", "none"), label = NULL)

tab_stat_rpct(data, total_label = NULL, total_statistic = "u_cases",
  total_row_position = c("below", "above", "none"), label = NULL)

tab_last_vstack(data, stat_position = c("outside_rows", "inside_rows"),
  stat_label = c("inside", "outside"), label = NULL)

tab_last_hstack(data, stat_position = c("outside_columns", "inside_columns"),
  stat_label = c("inside", "outside"), label = NULL)

tab_pivot(data, stat_position = c("outside_rows", "inside_rows",
```

```
  "outside_columns", "inside_columns"), stat_label = c("inside", "outside"))

tab_transpose(data)
```

Arguments

data	data.frame/intermediate_table
...	vector/data.frame/list. Variables for tables. Use mrset/mdset for multiple-response variables.
weight	numeric vector in tab_weight. Cases with NA's, negative and zero weights are removed before calculations.
total_row_position	Position of total row in the resulting table. Can be one of "below", "above", "none".
subgroup	logical vector in tab_subgroup. You can specify subgroup on which table will be computed.
label	character. Label for the statistic in the tab_stat_*.
unsafe	logical If TRUE than fun will be evaluated as is. It can lead to significant increase in the performance. But there are some limitations. For tab_stat_fun it means that your function fun should return vector of length one. Also there will be no attempts to make labels for statistic. For tab_stat_fun_df your function should return vector of length one or list/data.frame (optionally with 'row_labels' element - statistic labels). If unsafe is TRUE then further arguments (...) for fun will be ignored.
weighted_valid_n	logical. Sould we show weighted valid N in tab_stat_mean_sd_n? By default it is FALSE.
labels	character vector of length 3. Labels for mean, standard deviation and valid N in tab_stat_mean_sd_n.
total_label	By default "#Total". You can provide several names - each name for each total statistics.
total_statistic	By default it is "u_cases" (unweighted cases). Possible values are "u_cases", "u_responses", "u_cpct", "u_rpct", "u_tpct", "w_cases", "w_responses", "w_cpct", "w_rpct", "w_tpct". "u_" means unweighted statistics and "w_" means weighted statistics.
stat_position	character one of the values "outside_rows", "inside_rows", "outside_columns" or "inside_columns". It defines how we will combine statistics in the table.
stat_label	character one of the values "inside" or "outside". Where will be placed labels for the statistics relative to column names/row labels? See examples.

Details

- tab_cells variables on which percentage/cases/summary functions will be computed. Use mrset/mdset for multiple-response variables.

- tab_cols optional variables which breaks table by columns. Use mrset/mdset for multiple-response variables.
- tab_rows optional variables which breaks table by rows. Use mrset/mdset for multiple-response variables.
- tab_weight optional weight for the statistic.
- tab_mis_val optional missing values for the statistic. It will be applied on variables specified by tab_cells. It works in the same manner as na_if.
- tab_subgroup optional logical vector/expression which specify subset of data for table.
- tab_row_label Add to table empty row with specified row labels. It is usefull for making section headings and etc.
- tab_total_row_position Default value for total_row_position argument in tab_stat_cases and etc. Can be one of "below", "above", "none".
- tab_total_label Default value for total_label argument in tab_stat_cases and etc. You can provide several names - each name for each total statistics.
- tab_total_statistic Default value for total_statistic argument in tab_stat_cases and etc. You can provide several values. Possible values are "u_cases", "u_responses", "u_cpct", "u_rpct", "u_tpct", "w_cases", "w_responses", "w_cpct", "w_rpct", "w_tpct". "u_" means unweighted statistics and "w_" means weighted statistics.
- tab_stat_fun, tab_stat_fun_df tab_stat_fun applies function on each variable in cells separately, tab_stat_fun_df gives to function each data.frame in cells as a whole data.table with all names converted to variable labels (if labels exists). So it is not recommended to rely on original variables names in your fun. For details see cro_fun. You can provide several functions as arguments. They will be combined as with combine_functions. So you can use method argument. For details see documentation for combine_functions.
- tab_stat_cases calculate counts.
- tab_stat_cpct, tab_stat_cpct_responses calculate column percent. These functions give different results only for multiple response variables. For tab_stat_cpct base of percent is number of valid cases. Case is considered as valid if it has at least one non-NA value. So for multiple response variables sum of percent may be greater than 100. For tab_stat_cpct_responses base of percent is number of valid responses. Multiple response variables can have several responses for single case. Sum of percent of tab_stat_cpct_responses always equals to 100%.
- tab_stat_rpct calculate row percent. Base for percent is number of valid cases.
- tab_stat_tpct calculate table percent. Base for percent is number of valid cases.
- tab_stat_mean, tab_stat_median, tab_stat_se, tab_stat_sum, tab_stat_min, tab_stat_max, tab_stat_sd, tab_stat_valid_n, tab_stat_unweighted_valid_n different summary statistics. NA's are always omitted.
- tab_pivot finalize table creation and define how different tab_stat_* will be combined
- tab_transpose transpose final table after tab_pivot or last statistic.

Value

All of these functions return object of class intermediate_table except tab_pivot which returns final result - object of class etable. Basically it's a data.frame but class is needed for custom methods.

See Also

fre, cro, cro_fun, tab_sort_asc, drop_empty_rows, significance.

Examples

```
data(mtcars)
mtcars = apply_labels(mtcars,
                      mpg = "Miles/(US) gallon",
                      cyl = "Number of cylinders",
                      disp = "Displacement (cu.in.)",
                      hp = "Gross horsepower",
                      drat = "Rear axle ratio",
                      wt = "Weight (1000 lbs)",
                      qsec = "1/4 mile time",
                      vs = "Engine",
                      vs = c("V-engine" = 0,
                             "Straight engine" = 1),
                      am = "Transmission",
                      am = c("Automatic" = 0,
                             "Manual"=1),
                      gear = "Number of forward gears",
                      carb = "Number of carburetors"
)
# some examples from 'cro'
# simple example - generally with 'cro' it can be made with less typing
mtcars %>%
    tab_cells(cyl) %>%
    tab_cols(vs) %>%
    tab_stat_cpct() %>%
    tab_pivot()

# split rows
mtcars %>%
    tab_cells(cyl) %>%
    tab_cols(vs) %>%
    tab_rows(am) %>%
    tab_stat_cpct() %>%
    tab_pivot()

# multiple banners
mtcars %>%
    tab_cells(cyl) %>%
    tab_cols(total(), vs, am) %>%
    tab_stat_cpct() %>%
    tab_pivot()

# nested banners
mtcars %>%
    tab_cells(cyl) %>%
    tab_cols(total(), vs %nest% am) %>%
    tab_stat_cpct() %>%
    tab_pivot()
```

```
# summary statistics
mtcars %>%
    tab_cells(mpg, disp, hp, wt, qsec) %>%
    tab_cols(am) %>%
    tab_stat_fun(Mean = w_mean, "Std. dev." = w_sd, "Valid N" = w_n) %>%
    tab_pivot()

# summary statistics - labels in columns
mtcars %>%
    tab_cells(mpg, disp, hp, wt, qsec) %>%
    tab_cols(am) %>%
    tab_stat_fun(Mean = w_mean, "Std. dev." = w_sd, "Valid N" = w_n, method = list) %>%
    tab_pivot()

# subgroup with droping empty columns
mtcars %>%
    tab_subgroup(am == 0) %>%
    tab_cells(cyl) %>%
    tab_cols(total(), vs %nest% am) %>%
    tab_stat_cpct() %>%
    tab_pivot() %>%
    drop_empty_columns()

# total position at the top of the table
mtcars %>%
    tab_cells(cyl) %>%
    tab_cols(total(), vs) %>%
    tab_rows(am) %>%
    tab_stat_cpct(total_row_position = "above",
                  total_label = c("number of cases", "row %"),
                  total_statistic = c("u_cases", "u_rpct")) %>%
    tab_pivot()

# this example cannot be made easily with 'cro'
mtcars %>%
    tab_cells(am) %>%
    tab_cols(total(), vs) %>%
    tab_total_row_position("none") %>%
    tab_stat_cpct(label = "col %") %>%
    tab_stat_rpct(label = "row %") %>%
    tab_stat_tpct(label = "table %") %>%
    tab_pivot(stat_position = "inside_rows")

# statistic labels inside columns
mtcars %>%
    tab_cells(am) %>%
    tab_cols(total(), vs) %>%
    tab_total_row_position("none") %>%
    tab_stat_cpct(label = "col %") %>%
    tab_stat_rpct(label = "row %") %>%
    tab_stat_tpct(label = "table %") %>%
    tab_pivot(stat_position = "inside_columns")
```

```
# stacked statistics
mtcars %>%
    tab_cells(cyl) %>%
    tab_cols(total(), am) %>%
    tab_stat_mean() %>%
    tab_stat_se() %>%
    tab_stat_valid_n() %>%
    tab_stat_cpct() %>%
    tab_pivot()

# stacked statistics with section headings
mtcars %>%
    tab_cells(cyl) %>%
    tab_cols(total(), am) %>%
    tab_row_label("#Summary statistics") %>%
    tab_stat_mean() %>%
    tab_stat_se() %>%
    tab_stat_valid_n() %>%
    tab_row_label("#Column percent") %>%
    tab_stat_cpct() %>%
    tab_pivot()

# stacked statistics with different variables
mtcars %>%
    tab_cols(total(), am) %>%
    tab_cells(mpg, hp, qsec) %>%
    tab_stat_mean() %>%
    tab_cells(cyl, carb) %>%
    tab_stat_cpct() %>%
    tab_pivot()

# stacked statistics - label position outside row labels
mtcars %>%
    tab_cells(cyl) %>%
    tab_cols(total(), am) %>%
    tab_stat_mean() %>%
    tab_stat_se %>%
    tab_stat_valid_n() %>%
    tab_stat_cpct(label = "Col %") %>%
    tab_pivot(stat_label = "outside")

# example from 'cro_fun_df' - linear regression by groups with sorting
mtcars %>%
    tab_cells(dtfrm(mpg, disp, hp, wt, qsec)) %>%
    tab_cols(total(), am) %>%
    tab_stat_fun_df(
        function(x){
            frm = reformulate(".", response = names(x)[1])
            model = lm(frm, data = x)
            dtfrm('Coef. estimate' = coef(model),
                  confint(model)
            )
```

```
        }
    ) %>%
    tab_pivot() %>%
    tab_sort_desc()

# multiple-response variables and weight
data(product_test)
codeframe_likes = num_lab("
                          1 Liked everything
                          2 Disliked everything
                          3 Chocolate
                          4 Appearance
                          5 Taste
                          6 Stuffing
                          7 Nuts
                          8 Consistency
                          98 Other
                          99 Hard to answer
                          ")

set.seed(1)
product_test = compute(product_test, {
    # recode age by groups
    age_cat = recode(s2a, lo %thru% 25 ~ 1, lo %thru% hi ~ 2)

    var_lab(age_cat) = "Age"
    val_lab(age_cat) = c("18 - 25" = 1, "26 - 35" = 2)

    var_lab(a1_1) = "Likes. VSX123"
    var_lab(b1_1) = "Likes. SDF456"
    val_lab(a1_1) = codeframe_likes
    val_lab(b1_1) = codeframe_likes

    wgt = runif(.N, 0.25, 4)
    wgt = wgt/sum(wgt)*.N
})

product_test %>%
    tab_cells(mrset(a1_1 %to% a1_6), mrset(b1_1 %to% b1_6)) %>%
    tab_cols(total(), age_cat) %>%
    tab_weight(wgt) %>%
    tab_stat_cpct() %>%
    tab_sort_desc() %>%
    tab_pivot()

# trick to place cell variables labels inside columns
# useful to compare two variables
# '|' is needed to prevent automatic labels creation from argument
# alternatively we can use list(...) to avoid this
product_test %>%
    tab_cols(total(), age_cat) %>%
    tab_weight(wgt) %>%
    tab_cells("|" = unvr(mrset(a1_1 %to% a1_6))) %>%
```

```
    tab_stat_cpct(label = var_lab(a1_1)) %>%
    tab_cells("|" = unvr(mrset(b1_1 %to% b1_6))) %>%
    tab_stat_cpct(label = var_lab(b1_1)) %>%
    tab_pivot(stat_position = "inside_columns")

# if you need standard evaluation, use 'vars'
tables = mtcars %>%
      tab_cols(total(), am %nest% vs)

for(each in c("mpg", "disp", "hp", "qsec")){
    tables = tables %>% tab_cells(vars(each)) %>%
        tab_stat_fun(Mean = w_mean, "Std. dev." = w_sd, "Valid N" = w_n)
}
tables %>% tab_pivot()
```

`tab_significance_options` *Mark significant differences between columns of the table*

Description

- `significance_cpct` conducts z-tests between column percent in the result of cro_cpct. Results are calculated with the same formula as in prop.test without continuity correction.
- `significance_means` conducts t-tests between column means in the result of cro_mean_sd_n. Results are calculated with the same formula as in t.test.
- `significance_cases` conducts chi-squared tests on the subtable of table with counts in the result of cro_cases. Results are calculated with the same formula as in chisq.test without continuity correction.

There are four type of comparisons which can be conducted simultaneously (argument `compare_type`). `subtable` provide comparison between all columns inside each subtable. `previous_column` is comparison of each column in the subtable with previous column. It is useful if columns are periods or waves of survey. `first_column` provide comparison of table first column with all other columns in the table. `adjusted_first_column` is comparison with first column but with adjustment for common base. It is useful if first column is total column and other columns are subgroup of this total. Adjustments are made according to algorithm in IBM SPSS Statistics Algorithms v20, p. 263. Note that with these adjustments t-tests between means are made with equal variance assumed (as with `var_equal = TRUE`). By now there are no adjustments for multiple-response variables (results of mrset) in the table columns so significance tests are rather approximate for such cases. Also there are functions for significance testing in the sequence of custom tables calculations (see tables).

- `tab_last_sig_cpct`, `tab_last_sig_means` and `tab_last_sig_cpct` make the same tests as there analogs mentioned above. It is recommended to use them after appropriate statistic function: tab_stat_cpct, tab_stat_mean_sd_n and tab_stat_cases.
- `tab_significance_options` With this function we can set significance options for entire custom table creation sequence.

- tab_last_add_sig_labels This function applies add_sig_labels to last calculated table - it add labels (letters by default) for significance to columns header. It may be useful if you want to combine table with significance with table without it.
- tab_last_round This function rounds numeric columns in the last calculated table to specified number of digits. It is sometimes needed if you want to combine table with significance with table without it.

Usage

```
tab_significance_options(data, sig_level = 0.05, min_base = 2,
  delta_cpct = 0, delta_means = 0, compare_type = "subtable",
  bonferroni = FALSE, subtable_marks = "greater", inequality_sign = "both"
  %in% subtable_marks, sig_labels = LETTERS,
  sig_labels_previous_column = c("v", "^"), sig_labels_first_column = c("-",
  "+"), keep = c("percent", "cases", "means", "sd", "bases"),
  total_marker = "#", total_row = 1, digits = get_expss_digits(),
  na_as_zero = FALSE, var_equal = FALSE, mode = c("replace", "append"))

tab_last_sig_cpct(data, sig_level = 0.05, delta_cpct = 0, min_base = 2,
  compare_type = "subtable", bonferroni = FALSE,
  subtable_marks = c("greater", "both", "less"), inequality_sign = "both"
  %in% subtable_marks, sig_labels = LETTERS,
  sig_labels_previous_column = c("v", "^"), sig_labels_first_column = c("-",
  "+"), keep = c("percent", "bases"), na_as_zero = FALSE,
  total_marker = "#", total_row = 1, digits = get_expss_digits(),
  mode = c("replace", "append"), label = NULL)

tab_last_sig_means(data, sig_level = 0.05, delta_means = 0, min_base = 2,
  compare_type = "subtable", bonferroni = FALSE,
  subtable_marks = c("greater", "both", "less"), inequality_sign = "both"
  %in% subtable_marks, sig_labels = LETTERS,
  sig_labels_previous_column = c("v", "^"), sig_labels_first_column = c("-",
  "+"), keep = c("means", "sd", "bases"), var_equal = FALSE,
  digits = get_expss_digits(), mode = c("replace", "append"),
  label = NULL)

tab_last_sig_cases(data, sig_level = 0.05, min_base = 2, keep = c("cases",
  "bases"), total_marker = "#", total_row = 1,
  digits = get_expss_digits(), mode = c("replace", "append"),
  label = NULL)

tab_last_round(data, digits = get_expss_digits())

tab_last_add_sig_labels(data, sig_labels = LETTERS)

significance_cases(x, sig_level = 0.05, min_base = 2, keep = c("cases",
  "bases"), total_marker = "#", total_row = 1,
  digits = get_expss_digits())
```

```
significance_cpct(x, sig_level = 0.05, delta_cpct = 0, min_base = 2,
  compare_type = "subtable", bonferroni = FALSE,
  subtable_marks = c("greater", "both", "less"), inequality_sign = "both"
  %in% subtable_marks, sig_labels = LETTERS,
  sig_labels_previous_column = c("v", "^"), sig_labels_first_column = c("-",
  "+"), keep = c("percent", "bases"), na_as_zero = FALSE,
  total_marker = "#", total_row = 1, digits = get_expss_digits())

add_sig_labels(x, sig_labels = LETTERS)

significance_means(x, sig_level = 0.05, delta_means = 0, min_base = 2,
  compare_type = "subtable", bonferroni = FALSE,
  subtable_marks = c("greater", "both", "less"), inequality_sign = "both"
  %in% subtable_marks, sig_labels = LETTERS,
  sig_labels_previous_column = c("v", "^"), sig_labels_first_column = c("-",
  "+"), keep = c("means", "sd", "bases"), var_equal = FALSE,
  digits = get_expss_digits())
```

Arguments

data	data.frame/intermediate_table for tab_* functions.
sig_level	numeric. Significance level - by default it equals to 0.05.
min_base	numeric. Significance test will be conducted if both columns have bases greater than min_base. By default it equals to 2.
delta_cpct	numeric. Minimal delta between percent for which we mark significant differences (in percent points) - by default it equals to zero. Note that, for example, for minimal 5 percent difference delta_cpct should be equals 5, not 0.05.
delta_means	numeric. Minimal delta between means for which we mark significant differences - by default it equals to zero.
compare_type	Type of compare between columns. By default it is subtable - comparisons will be conducted between columns of each subtable. Other possible values are: first_column, adjusted_first_column and previous_column. We can conduct several tests simultaneously.
bonferroni	logical. FALSE by default. Should we use Bonferroni adjustment by number of comparisons in each row?
subtable_marks	character. One of "greater", "both" or "less". By deafult we mark only values which are significantly greater than some other columns. We can change this behavior by setting argument to less or both.
inequality_sign	logical. FALSE if subtable_marks is "less" or "greater". Should we show > or < before significance marks of subtable comparisons.
sig_labels	character vector. Labels for marking differences between columns of subtable.
sig_labels_previous_column	a character vector with two elements. Labels for marking difference with previous column. First mark means 'lower' (by default it is v) and the second means greater (^).

sig_labels_first_column
: a character vector with two elements. Labels for marking difference with first column of the table. First mark means 'lower' (by default it is -) and the second means 'greater' (+).

keep
: character. One or more from "percent", "cases", "means", "bases", "sd" or "none". This argument determines which statistics will remain in the table after significance marking.

total_marker
: character. Mark of total rows in table.

total_row
: integer/character. In case of several totals per subtable it is number or name of total row for significance calculation.

digits
: an integer indicating how much digits after decimal separator will be shown in the final table.

na_as_zero
: logical. FALSE by default. Should we treat NA's as zero cases?

var_equal
: a logical variable indicating whether to treat the two variances as being equal. For details see t.test.

mode
: character. One of replace(default) or append. In the first case the previous result in the sequence of table calculation will be replaced with result of significance testing. In the second case result of the significance testing will be appended to sequence of table calculation.

label
: character. Label for the statistic in the tab_*. Ignored if the mode is equals to replace.

x
: table (class etable): result of cro_cpct with proportions and bases for significance_cpct, result of cro_mean_sd_n with means, standard deviations and valid N for significance_means, and result of cro_cases with counts and bases for significance_cases.

Value

tab_last_* functions return objects of class intermediate_table. Use tab_pivot to get final result - object of class etable. Other functions return object of class etable with marks of significant differences between columns.

See Also

cro_cpct, cro_cases, cro_mean_sd_n, tables, compare_proportions, compare_means, prop.test, t.test, chisq.test

Examples

```
data(mtcars)
mtcars = apply_labels(mtcars,
                      mpg = "Miles/(US) gallon",
                      cyl = "Number of cylinders",
                      disp = "Displacement (cu.in.)",
                      hp = "Gross horsepower",
                      drat = "Rear axle ratio",
                      wt = "Weight (lb/1000)",
                      qsec = "1/4 mile time",
```

```
                      vs = "Engine",
                      vs = c("V-engine" = 0,
                             "Straight engine" = 1),
                      am = "Transmission",
                      am = c("Automatic" = 0,
                             "Manual"=1),
                      gear = "Number of forward gears",
                      carb = "Number of carburetors"
)

mtcars_table = calculate(mtcars,
                  cro_cpct(list(cyl, gear),
                           list(total(), vs, am))
                       )

significance_cpct(mtcars_table)

# comparison with first column
significance_cpct(mtcars_table, compare_type = "first_column")

# comparison with first column and inside subtable
significance_cpct(mtcars_table,
            compare_type = c("first_column", "subtable"))

# only significance marks
significance_cpct(mtcars_table, keep = "none")

# means
mtcars_means = calculate(mtcars,
                  cro_mean_sd_n(list(mpg, wt, hp),
                                list(total(), vs, cyl))
                       )

significance_means(mtcars_means)

# mark values which are less and greater
significance_means(mtcars_means, subtable_marks = "both")

# chi-squared test
mtcars_cases = calculate(mtcars,
                  cro_cases(list(cyl, gear),
                           list(total(), vs, am))
                       )

significance_cases(mtcars_cases)

# custom tables with significance
mtcars %>%
    tab_significance_options(subtable_marks = "both") %>%
    tab_cells(mpg, hp) %>%
    tab_cols(total(), vs, am) %>%
    tab_stat_mean_sd_n() %>%
    tab_last_sig_means(keep = "means") %>%
```

```
    tab_cells(cyl, gear) %>%
    tab_stat_cpct() %>%
    tab_last_sig_cpct() %>%
    tab_pivot()

# Overcomplicated examples - we move significance marks to
# separate columns. Columns with statistics remain numeric
mtcars %>%
    tab_significance_options(keep = "none",
                         sig_labels = NULL,
                         subtable_marks = "both",
                         mode = "append") %>%
    tab_cols(total(), vs, am) %>%
    tab_cells(mpg, hp) %>%
    tab_stat_mean_sd_n() %>%
    tab_last_sig_means() %>%
    tab_last_hstack("inside_columns") %>%
    tab_cells(cyl, gear) %>%
    tab_stat_cpct() %>%
    tab_last_sig_cpct() %>%
    tab_last_hstack("inside_columns") %>%
    tab_pivot(stat_position = "inside_rows") %>%
    drop_empty_columns()
```

`tab_sort_asc` *Partially (inside blocks) sort tables/data.frames*

Description

`tab_sort_asc`/`tab_sort_desc` sort tables (usually result of cro/tables) in ascending/descending order between specified rows (by default, it is rows which contain '#' in the first column).

Usage

```
tab_sort_asc(x, ..., excluded_rows = "#", na.last = FALSE)

tab_sort_desc(x, ..., excluded_rows = "#", na.last = TRUE)
```

Arguments

x	data.frame
...	character/numeric or criteria/logical functions (see criteria). Column names/numbers for data.frame/matrix by which object will be sorted. Names at the top-level can be unquoted (non-standard evaluation). For standard evaluation of parameters you can surround them by round brackets. See examples. If this argument is missing then table will be sorted by second column. Usually second column is the first column with numbers in the table (there are row labels in the first column).

excluded_rows	character/logical/numeric rows which won't be sorted. Rows of the table will be sorted between excluded rows. If it is characters then they will be considered as pattern/vector of patterns. Patterns will be matched with Perl-style regular expression with values in the first column of x (see grep, perl = TRUE argument). Rows which have such patterns will be excluded. By default, pattern is "#" because "#" marks totals in the result of cro.
na.last	for controlling the treatment of NAs. If TRUE, missing values in the data are put last; if FALSE, they are put first; if NA, they are removed.

Value

sorted table('etable')/data.frame

Examples

```
data(mtcars)

# apply labels
mtcars = apply_labels(mtcars,
    cyl = "Number of cylinders",
    vs = "Engine",
    vs = c("V-engine" = 0,
                   "Straight engine" = 1),
    am = "Transmission",
    am = c(automatic = 0,
                   manual=1),
    gear = "Number of forward gears",
    carb = "Number of carburetors"
)

# without sorting
mtcars %>% calculate(cro_cpct(list(cyl, gear, carb), list("#total", vs, am)))

# with sorting
mtcars %>%
    calculate(cro_cpct(list(cyl, gear, carb), list("#total", vs, am))) %>%
    tab_sort_desc

# sort by parameter
sorting_column = "Engine|V-engine"

mtcars %>%
    calculate(cro_cpct(list(cyl, gear, carb), list("#total", vs, am))) %>%
    tab_sort_desc((sorting_column))
```

---

unlab *Drop variable label and value labels*

Description

unlab returns variable x without variable labels and value labels

Usage

```
unlab(x)

drop_all_labels(x)
```

Arguments

x	Variable(s). Vector/data.frame/list.

Value

unlab returns original variable x without variable label, value labels and class.

See Also

unvr unvl

Examples

```
raw_var = rep(1:2,5)
var_with_lab = set_var_lab(raw_var,"Income")
val_lab(var_with_lab) = c("Low"=1,"High"=2)
identical(raw_var,unlab(var_with_lab)) # should be TRUE
```

---

values2labels *Replace vector/matrix/data.frame/list values with corresponding value labels.*

Description

values2labels replaces vector/matrix/data.frame/list values with corresponding value labels. If there are no labels for some values they are converted to characters in most cases. If there are no labels at all for variable it remains unchanged. v2l is just shortcut to values2labels.

Usage

```
values2labels(x)

v2l(x)
```

Arguments

x	vector/matrix/data.frame/list

Value

Object of the same form as x but with value labels instead of values.

See Also

names2labels, val_lab, var_lab

Examples

```
data(mtcars)
mtcars = modify(mtcars,{
                var_lab(mpg) = NULL
                val_lab(am) = c(" automatic" = 0, " manual" =  1)
})

summary(lm(mpg ~ ., data = values2labels(mtcars[,c("mpg","am")])))
```

---

val_lab *Set or get value labels*

Description

These functions set/get/drop value labels. Duplicated values are not allowed. If argument x is data.frame or list then labels applied to all elements of data.frame/list. To drop value labels, use `val_lab(var) <-NULL` or `unvl(var)`. `make_labels` converts text from the form that usually used in questionnaires to named vector. For variable labels see var_lab. For working with entire data.frame see apply_labels.

- `val_lab` returns value labels or NULL if labels doesn't exist.
- `val_lab<-` set value labels.
- `set_val_lab` returns variable with value labels.
- `add_val_lab<-` add value labels to already existing value labels.
- `unvl` drops value labels.
- `make_labels` makes named vector from text for usage as value labels.
- `num_lab`, `lab_num` and `autonum` are shortcuts for `make_labels` with `code_postion` 'left', 'right' and 'autonum' accordingly.

Usage

```
val_lab(x)

val_lab(x) <- value

set_val_lab(x, value, add = FALSE)

add_val_lab(x, value)
```

```
add_val_lab(x) <- value

unvl(x)

drop_val_labs(x)

make_labels(text, code_position = c("left", "right", "autonum"))

drop_unused_labels(x)

num_lab(text)

lab_num(text)

autonum(text)
```

Arguments

x	Variable(s). Vector/data.frame/list.
value	Named vector. Names of vector are labels for the appropriate values of variable x.
add	Logical. Should we add value labels to old labels or replace it? Deafult is FALSE - we completely replace old values. If TRUE new value labels will be combined with old value labels.
text	text that should be converted to named vector
code_position	Possible values "left", "right" - position of numeric code in text. "autonum" - makes codes by autonumbering lines of text.

Details

Value labels are stored in attribute "labels" (attr(x,"labels")). We set variable class to "labelled" for preserving labels from dropping during some operations (such as c and `[`).

Value

val_lab return value labels (named vector). If labels doesn't exist it return NULL . val_lab<- and set_val_lab return variable (vector x) of class "labelled" with attribute "labels" which contains value labels. make_labels return named vector for usage as value labels.

Examples

```
# toy example
set.seed(123)
# score - evaluation of tested product

score = sample(-1:1,20,replace = TRUE)
var_lab(score) = "Evaluation of tested brand"
val_lab(score) = c("Dislike it" = -1,
```

```
                    "So-so" = 0,
                    "Like it" = 1
                    )

# frequency of product scores
fre(score)

# brands - multiple response question
# Which brands do you use during last three months?

brands = as.dtfrm(t(replicate(20,sample(c(1:5,NA),4,replace = FALSE))))

var_lab(brands) = "Used brands"
val_lab(brands) = make_labels("
                              1 Brand A
                              2 Brand B
                              3 Brand C
                              4 Brand D
                              5 Brand E
                              ")


# percentage of used brands
fre(brands)

# percentage of brands within each score
cro_cpct(brands, score)

## make labels from text copied from questionnaire

age = c(1, 2, 1, 2)

val_lab(age) = num_lab("
 1. 18 - 26
 2. 27 - 35
")

# note support of value lables in base R
table(age)

# or, if in original codes is on the right side

products = 1:8

val_lab(products) = lab_num("
 Chocolate bars    1
 Chocolate sweets (bulk) 2
 Slab chocolate(packed) 3
 Slab chocolate (bulk) 4
 Boxed chocolate sweets 5
 Marshmallow/pastilles in chocolate coating 6
 Marmalade in chocolate coating 7
 Other 8
```

```
")

table(products)
```

`vars` *Get variables/range of variables by name/by pattern.*

Description

- `vars` returns data.frame with all variables by their names or by criteria (see criteria). There is no non-standard evaluation in this function by design so use quotes for names of your variables. This function is intended to get variables by parameter/criteria. The only exception with non-standard evaluation is `%to%`. You can use `%to%` inside `vars` or independently.
- `..[]` returns data.frame with all variables by their names or by criteria (see criteria). Names at the top-level can be unquoted (non-standard evaluation). For standard evaluation of parameters you can surround them by round brackets. You can assign to this expression. If there are several names inside square brackets then each element of list/data.frame from right side will be assigned to appropriate name from left side. You can use `item1 %to% item2` notation to get/create sequence of variables. If there are no arguments inside square brackets than from each item of RHS will be created separate variable in the parent frame. In this case RHS should be named list or data.frame.
- `..$name` sets/returns object which name is stored in the variable `name`. It is convenient wrapper around get/assign functions.
- `%to%` returns range of variables between `e1` and `e2` (similar to SPSS 'to'). modify, modify_if, calculate, keep, except and where support `%to%`. Inside global environment with, within `%to%` will take range from names of variables sorted in the alphabetic order.
- `indirect`/`indirect_list` are aliases for `vars`/`vars_list`.

Functions with word 'list' in name return lists of variables instead of dataframes.

Usage

```
vars(...)

vars_list(...)

indirect(...)

indirect_list(...)

e1 %to% e2

e1 %to_list% e2

..
```

Arguments

...	characters names of variables or criteria/logical functions
e1	unquoted name of start variable (e. g. a_1)
e2	unquoted name of start variable (e. g. a_5)

Format

An object of class `parameter` of length 1.

Value

data.frame/list with variables

See Also

keep, except, do_repeat, compute, calculate, where

Examples

```
# In data.frame
dfs = data.frame(
    a = rep(1, 5),
    b_1 = rep(11, 5),
    b_2 = rep(12, 5),
    b_3 = rep(13, 5),
    b_4 = rep(14, 5),
    b_5 = rep(15, 5)
)

# calculate sum of b_* variables
compute(dfs, {
    b_total = sum_row(b_1 %to% b_5)
})

# In global environement
a = rep(10, 5)
a1 = rep(1, 5)
a2 = rep(2, 5)
a3 = rep(3, 5)
a4 = rep(4, 5)
a5 = rep(5, 5)

# identical results
a1 %to% a5
vars(perl("^a[0-9]$"))
..[perl("^a[0-9]$")]

# sum each row
sum_row(a1 %to% a5)

# variable substitution
```

```
name1 = "a"
name2 = "new_var"

# in global environment
..$name1 # give as variable 'a'

..$name2 = ..$name1 * 2 # create variable 'new_var' which is equal to 'a' times 2
new_var

# inside data.frame
compute(dfs, {
     ..$name2 = ..$name1*2
})

compute(dfs, {
     for(name1 in paste0("b_", 1:5)){
         name2 = paste0("new_", name1)
         ..$name2 = ..$name1*2
     }
     rm(name1, name2) # we don't need this variables as columns in 'dfs'
})

# square brackets notation - multi-assignment
name1 = paste0("b_", 1:5)
compute(dfs, {
          # round brackets about 'name1' is needed to avoid using it 'as is'
         ..[paste0("new_", name1)] = ..[(name1)]*2
})

# the same result
# note the automatic creation of sequence of variables
compute(dfs, {
         ..[new_b_1 %to% new_b_5] = ..[b_1 %to% b_5]*2
})

# assignment form of 'recode' on multiple variables
compute(dfs, {
         recode(..[b_1 %to% b_5]) = 13 %thru% hi ~ 20
})

# empty brackets - unboxing of dichotomy.
compute(dfs, {
         ..[] =  as.dichotomy(b_1 %to% b_5, prefix = "v_")
})
```

---

var_lab *Set or get variable label*

Description

These functions set/get/drop variable labels. For value labels see val_lab. For working with entire data.frame see apply_labels.

- `var_lab` returns variable label or NULL if label doesn't exist.
- `var_lab<-` set variable label.
- `set_var_lab` returns variable with label.
- `unvr` drops variable label.

Usage

```
var_lab(x)

var_lab(x) <- value

set_var_lab(x, value)

unvr(x)

drop_var_labs(x)
```

Arguments

x	Variable. In the most cases it is numeric vector.
value	A character scalar - label for the variable x.

Details

Variable label is stored in attribute "label" (`attr(x,"label")`). For preserving from dropping this attribute during some operations (such as `c`) variable class is set to "labelled". There are special methods of subsetting and concatenation for this class. To drop variable label use `var_lab(var) <- NULL` or `unvr(var)`.

Value

`var_lab` return variable label. If label doesn't exist it return NULL . `var_lab<-` and `set_var_lab` return variable (vector x) of class "labelled" with attribute "label" which equals submitted value.

Examples

```
data(mtcars)
mtcars = modify(mtcars,{
                var_lab(mpg) = "Miles/(US) gallon"
                var_lab(cyl) = "Number of cylinders"
                var_lab(disp) = "Displacement (cu.in.)"
                var_lab(hp) = "Gross horsepower"
                var_lab(drat) = "Rear axle ratio"
                var_lab(wt) = "Weight (lb/1000)"
                var_lab(qsec) = "1/4 mile time"
```

```
                var_lab(vs) = "V/S"
                var_lab(am) = "Transmission (0 = automatic, 1 = manual)"
                val_lab(am) = c(automatic = 0, manual=1)
                var_lab(gear) = "Number of forward gears"
                var_lab(carb) = "Number of carburetors"
})

fre(mtcars$am)

calculate(mtcars,
     cro_mean(list(mpg, disp, hp, qsec), list(total(), am))
     )
```

`vectors` *Infix operations on vectors - append, diff, intersection, union, replication*

Description

- `%a%` a(ppends) second argument to first argument. See also append.
- `%u%` and `v_union` u(nite) first and second arguments. Remove elements from second argument which exist in first argument.
- `%d%` and `v_diff` d(iff) second argument from first argument. Second argument could be a function which returns logical value. In this case elements of first argument which give TRUE will be removed.
- `%i%` and `v_intersect` i(ntersect) first argument and second argument. Second argument could be a function which returns logical value. In this case elements of first argument which give FALSE will be removed.
- `%e%` and `v_xor` e(xclusive OR). Returns elements that contained only in one of arguments.
- `%r%` r(epeats) first argument second argument times. See also rep.
- `%n_d%` and `n_diff` n(ames) d(iff) - diff second argument from names of first argument. Second argument could be a function which returns logical value. In this case elements of first argument which names give TRUE will be removed.
- `%n_i%` and `n_intersect` n(ames) i(ntersect) - intersect names of first argument with second argument. Second argument could be a function which returns logical value. In this case elements of first argument which names give FALSE will be removed.

All these functions except `%n_d%`, `%n_i%` preserve names of vectors and don't remove duplicates. For `%d%`, `%i%`, `%n_d%`, `%n_i%` one can use criteria functions. See criteria for details.

Usage

```
e1 %a% e2

v_union(e1, e2)
```

```
e1 %u% e2

v_diff(e1, e2)

e1 %d% e2

v_intersect(e1, e2)

e1 %i% e2

v_xor(e1, e2)

e1 %e% e2

e1 %r% e2

n_intersect(e1, e2)

e1 %n_i% e2

n_diff(e1, e2)

e1 %n_d% e2
```

Arguments

e1	vector (possibly data.frame/matrix/list for %n_d%, %n_i%)
e2	vector (or function for %d%, %i%)

Value

vector (possibly data.frame/matrix/list for %n_d%, %n_i%)

Examples

```
1:4 %a% 5:6   # 1:6

1:4 %a% 4:5   # 1,2,3,4,4,5

1:4 %u% 4:5   # 1,2,3,4,5

1:6 %d% 5:6   # 1:4

# function as criterion
1:6 %d% gt(4) # 1:4

1:4 %i% 4:5   # 4
```

```
# function as criterion
letters %i% perl("[a-d]") # a,b,c,d

# function as criterion
letters %i% (fixed("a") | fixed("z")) # a, z

1:4 %e% 4:5   # 1, 2, 3, 5

1:2 %r% 2     # 1, 2, 1, 2

# %n_i%, %n_d%

# remove column Species
iris %n_d% "Species"

# leave only columns which names start with "Sepal"
iris %n_i% perl("^Sepal")

# leave column "Species" and columns which names start with "Sepal"
iris %n_i% (perl("^Sepal")|"Species")
```

---

vlookup *Look up values in dictionary.*

Description

vlookup/vlookup_df function is inspired by VLOOKUP spreadsheet function. It looks for a lookup_value in the lookup_column of the dict, and then returns values in the same rows from result_column. add_columns inspired by MATCH FILES (Add variables...) from SPSS Statistics. It works similar to SQL left join but number of cases in the left part always remain the same. If there are duplicated keys in the dict then error will be raised by default. .add_columns is the same function for default dataset.

Usage

```
vlookup(lookup_value, dict, result_column = 2, lookup_column = 1)

vlookup_df(lookup_value, dict, result_column = NULL, lookup_column = 1)

add_columns(data, dict, by = NULL, ignore_duplicates = FALSE)

.add_columns(dict, by = NULL, ignore_duplicates = FALSE)
```

Arguments

lookup_value	Vector of looked up values
dict	data.frame/matrix. Dictionary. Can be vector for vlookup/vlookup_df.

result_column numeric or character. Resulting columns of dict. There are special values: 'row.names', 'rownames', 'names'. If result_column equals to one of these special values and dict is matrix/data.frame then row names of dict will be returned. If dict is vector then names of vector will be returned. For vlookup_df default result_column is NULL and result will be entire rows. For vlookup defaut result_column is 2 - for frequent case of dictionary with keys in the first column and results in the second column.

lookup_column Column of dict in which lookup value will be searched. By default it is the first column of the dict. There are special values: 'row.names', 'rownames', 'names'. If lookup_column equals to one of these special values and dict is matrix/data.frame then values will be searched in the row names of dict. If dict is vector then values will be searched in names of the dict.

data data.frame to be joined with dict.

by character vector or NULL(default) or 1. Names of common variables in the data and dict by which we will attach dict to data. If it is NULL then common names will be used. If it is equals to 1 then we will use the first column from both dataframes. To add columns by different variables on data and dict use a named vector. For example, by = c("a" = "b") will match data.a to dict.b.

ignore_duplicates logical Should we ignore duplicates in the by variables in the dict? If it is TRUE than first occurence of duplicated key will be used.

Value

vlookup always return vector, vlookup_df always returns data.frame. row.names in result of vlookup_df are not preserved.

Examples

```
# with data.frame
dict = data.frame(num=1:26, small=letters, cap=LETTERS, stringsAsFactors = FALSE)
rownames(dict) = paste0('rows', 1:26)
identical(vlookup_df(1:3, dict), dict[1:3,]) # should be TRUE
vlookup(c(45,1:3,58), dict, result_column='cap')
vlookup_df(c('z','d','f'), dict, lookup_column = 'small')
vlookup_df(c('rows7', 'rows2', 'rows5'), dict, lookup_column = 'row.names')

# with vector
dict=1:26
names(dict) = letters

vlookup(c(2,4,6), dict, result_column='row.names')

# The same results
vlookup(c(2,4,6), dict, result_column='rownames')
vlookup(c(2,4,6), dict, result_column='names')

# example for 'add_columns' from base 'merge'
authors = sheet(
    surname = c("Tukey", "Venables", "Tierney", "Ripley", "McNeil"),
```

```
    nationality = c("US", "Australia", "US", "UK", "Australia"),
    deceased = c("yes", rep("no", 4))
)

books = sheet(
    surname = c("Tukey", "Venables", "Tierney",
                "Ripley", "Ripley", "McNeil", "R Core"),
    title = c("Exploratory Data Analysis",
              "Modern Applied Statistics ...",
              "LISP-STAT",
              "Spatial Statistics", "Stochastic Simulation",
              "Interactive Data Analysis",
              "An Introduction to R")
)

add_columns(books, authors)

# Just for fun. Examples borrowed from Microsoft Excel.
# It is not the R way of doing things.

# Example 2

ex2 = utils::read.table(header = TRUE, text = "
    Item_ID Item Cost Markup
    ST-340 Stroller 145.67  0.30
    BI-567 Bib 3.56  0.40
    DI-328 Diapers  21.45  0.35
    WI-989 Wipes  5.12  0.40
    AS-469 Aspirator 2.56  0.45
", stringsAsFactors = FALSE)

# Calculates the retail price of diapers by adding the markup percentage to the cost.
vlookup("DI-328", ex2, 3) * (1 + vlookup("DI-328", ex2, 4)) # 28.9575

# Calculates the sale price of wipes by subtracting a specified discount from
# the retail price.
(vlookup("WI-989", ex2, "Cost") * (1 + vlookup("WI-989", ex2, "Markup"))) * (1 - 0.2)  # 5.7344

A2 = ex2[1, "Item_ID"]
A3 = ex2[2, "Item_ID"]

# If the cost of an item is greater than or equal to $20.00, displays the string
# "Markup is nn%"; otherwise, displays the string "Cost is under $20.00".
ifelse(vlookup(A2, ex2, "Cost") >= 20,
       paste0("Markup is " , 100 * vlookup(A2, ex2, "Markup"),"%"),
       "Cost is under $20.00") # Markup is 30%


# If the cost of an item is greater than or equal to $20.00, displays the string
# Markup is nn%"; otherwise, displays the string "Cost is $n.nn".
ifelse(vlookup(A3, ex2, "Cost") >= 20,
       paste0("Markup is: " , 100 * vlookup(A3, ex2, "Markup") , "%"),
       paste0("Cost is $", vlookup(A3, ex2, "Cost"))) #Cost is $3.56
```

```
# Example 3

ex3 = utils::read.table(header = TRUE, text = "
    ID  Last_name  First_name  Title Birth_date
    1 Davis Sara 'Sales Rep.'  12/8/1968
    2 Fontana Olivier 'V.P. of Sales' 2/19/1952
    3 Leal Karina 'Sales Rep.' 8/30/1963
    4 Patten Michael 'Sales Rep.' 9/19/1958
    5 Burke Brian 'Sales Mgr.' 3/4/1955
    6 Sousa Luis 'Sales Rep.'  7/2/1963
", stringsAsFactors = FALSE)

# If there is an employee with an ID of 5, displays the employee's last name;
# otherwise, displays the message "Employee not found".
if_na(vlookup(5, ex3, "Last_name"), "Employee not found") # Burke

# Many employees
if_na(vlookup(1:10, ex3, "Last_name"), "Employee not found")

# For the employee with an ID of 4, concatenates the values of three cells into
# a complete sentence.
paste0(vlookup(4, ex3, "First_name"), " ",
       vlookup(4, ex3, "Last_name"), " is a ",
       vlookup(4, ex3, "Title")) # Michael Patten is a Sales Rep.
```

`where` *Subset (filter) data.frames/matrices/vectors/lists*

Description

For the data frame `cond` will be evaluated in the data.frame's context. So columns can be referred as variables in the expression (see the examples). If `data` is list then `where` will be applied to each element of the list. For other types (vector/matrix) there is no non-standard evaluation. There is a special constant `.N` which equals to number of rows in `data` for usage in `cond` expression. `.where` is version for working with default dataset. See default_dataset.

Usage

```
where(data, cond)

.where(cond)
```

Arguments

`data`	data.frame/matrix/vector/list to be subsetted
`cond`	logical or numeric expression indicating elements or rows to keep: missing values (NA) are taken as FALSE. If `data` is data.frame then `cond` will be evaluated in the scope of the `data`.

Value

data.frame/matrix/vector/list which contains just selected rows.

Examples

```
# leave only 'setosa'
where(iris, Species == "setosa")
# leave only first five rows
where(iris, 1:5)


# example of .N usage.
set.seed(42)
train = where(iris, sample(.N, 100))
str(train)

set.seed(42)
test = where(iris, -sample(.N, 100))
str(test)

# list example
set.seed(123)
rand_matr = matrix(sample(10, 60, replace = TRUE), ncol = 3)
rand_vec = sample(10, 20, replace = TRUE)
my_list = list(iris, rand_matr, rand_vec)
# two random elements from the each list item
where(my_list, sample(.N, 2))
```

---

window_fun *Function over grouping variables (window function)*

Description

This is faster version of ave. window_fun applies function to every subset of x and return vector of the same length as x.

Usage

```
window_fun(x, ...)
```

Arguments

x	A vector
...	Grouping variables all of the same length as x or length 1 and function as last argument.

Value

vector of the same length as x

Examples

```
window_fun(1:3, mean)  # no grouping -> grand mean

attach(warpbreaks)

window_fun(breaks, wool, mean)
window_fun(breaks, tension, function(x) mean(x, trim = 0.1))

detach(warpbreaks)
```

`write_labels` *Write data with labels to file in R code or in SPSS syntax.*

Description

`write_labelled_*` functions write data in the CSV format and file with R code/SPSS syntax for labelling data. SPSS syntax also contains code for reading data in SPSS. `write_labelled_*` doesn't save rownames of data.frame. `write_labels_*` functions write R code/SPSS syntax for labelling data. It allows to extract labels from *.sav files that come without accompanying syntax. `read_labelled_csv` reads data file in CSV format and apply labels from accompanying file with R code. `*_csv2` write/read data with semicolon separator and comma as decimal delimiter. `*_tab`/`*_tab2` write/read data with tab separator and "."/"," as decimal delimiter.

Usage

```
write_labels(x, filename, fileEncoding = "")

read_labelled_csv(filename, fileEncoding = "", encoding = "unknown",
  sep = ",", dec = ".", undouble_quotes = TRUE, ...)

read_labelled_csv2(filename, fileEncoding = "", encoding = "unknown",
  sep = ";", dec = ",", undouble_quotes = TRUE, ...)

read_labelled_tab(filename, fileEncoding = "", encoding = "unknown",
  sep = "\t", dec = ".", undouble_quotes = TRUE, ...)

read_labelled_tab2(filename, fileEncoding = "", encoding = "unknown",
  sep = "\t", dec = ",", undouble_quotes = TRUE, ...)

write_labelled_csv(x, filename, fileEncoding = "", sep = ",", dec = ".",
  qmethod = c("double", "escape"), remove_new_lines = TRUE, ...)

write_labelled_csv2(x, filename, fileEncoding = "", sep = ";", dec = ",",
  qmethod = c("double", "escape"), remove_new_lines = TRUE, ...)

write_labelled_tab(x, filename, fileEncoding = "", sep = "\t", dec = ".",
  qmethod = c("double", "escape"), remove_new_lines = TRUE, ...)
```

```
write_labelled_tab2(x, filename, fileEncoding = "", sep = "\t",
  dec = ",", qmethod = c("double", "escape"), remove_new_lines = TRUE,
  ...)

write_labelled_spss(x, filename, fileEncoding = "", remove_new_lines = TRUE,
  ...)

write_labels_spss(x, filename)
```

Arguments

x	data.frame to be written/data.frame whose labels to be written
filename	the name of the file which the data are to be read from/write to.
fileEncoding	character string: if non-empty declares the encoding to be used on a file (not a connection) so the character data can be re-encoded as they are written. Used for writing dictionary. See file.
encoding	default is "unknown". Other possible options are "UTF-8" and "Latin-1". Note: it is not used to re-encode the input, rather enables handling of encoded strings in their native encoding. Used for writing data file. See fread.
sep	the field separator string. Values within each row of x are separated by this string.
dec	the string to use for decimal points in numeric or complex columns: must be a single character.
undouble_quotes	A logical indicating should we undouble quotes which were escaped by doubling (see qmethod). TRUE by default. Argument will be removed when data.table issue #1109 will be fixed.
...	additional arguments for fwrite/fread
qmethod	A character string specifying how to deal with embedded double quote characters when quoting strings. "escape" - the quote character (as well as the backslash character) is escaped in C style by a backslash, or "double" (default), in which case the double quote is doubled with another one.
remove_new_lines	A logical indicating should we replace new lines with spaces in the character variables. TRUE by default.

Value

Functions for writing invisibly return NULL. Functions for reading return labelled data.frame.

Examples

```
## Not run:
data(mtcars)
mtcars = modify(mtcars,{
               var_lab(mpg) = "Miles/(US) gallon"
```

```
                  var_lab(cyl) = "Number of cylinders"
                  var_lab(disp) = "Displacement (cu.in.)"
                  var_lab(hp) = "Gross horsepower"
                  var_lab(drat) = "Rear axle ratio"
                  var_lab(wt) = "Weight (lb/1000)"
                  var_lab(qsec) = "1/4 mile time"
                  var_lab(vs) = "Engine"
                  val_lab(vs) = c("V-engine" = 0,
                                  "Straight engine" = 1)
                  var_lab(am) = "Transmission"
                  val_lab(am) = c(automatic = 0,
                                  manual=1)
                  var_lab(gear) = "Number of forward gears"
                  var_lab(carb) = "Number of carburetors"
})

# to R code
# rownames are not preserved
write_labelled_csv(mtcars, "mtcars.csv")
new_mtcars = read_labelled_csv("mtcars.csv")

# to SPSS syntax
write_labelled_spss(mtcars, "mtcars.csv")


## End(Not run)
```

`w_mean` *Compute various weighted statistics*

Description

- `w_mean` weighted mean of a numeric vector
- `w_sd` weighted sample standard deviation of a numeric vector
- `w_var` weighted sample variance of a numeric vector
- `w_se` weighted standard error of a numeric vector
- `w_median` weighted median of a numeric vector
- `w_mad` weighted mean absolute deviation from median of a numeric vector
- `w_sum` weighted sum of a numeric vector
- `w_n` weighted number of values of a numeric vector
- `w_cov` weighted covariance matrix of a numeric matrix/data.frame
- `w_cor` weighted Pearson correlation matrix of a numeric matrix/data.frame
- `w_pearson` shortcut for `w_cor`. Weighted Pearson correlation matrix of a numeric matrix/data.frame
- `w_spearman` weighted Spearman correlation matrix of a numeric matrix/data.frame

Usage

```
w_mean(x, weight = NULL, na.rm = TRUE)

w_median(x, weight = NULL, na.rm = TRUE)

w_var(x, weight = NULL, na.rm = TRUE)

w_sd(x, weight = NULL, na.rm = TRUE)

w_se(x, weight = NULL, na.rm = TRUE)

w_mad(x, weight = NULL, na.rm = TRUE)

w_sum(x, weight = NULL, na.rm = TRUE)

w_n(x, weight = NULL, na.rm = TRUE)

unweighted_valid_n(x, weight = NULL)

valid_n(x, weight = NULL)

w_max(x, weight = NULL, na.rm = TRUE)

w_min(x, weight = NULL, na.rm = TRUE)

w_cov(x, weight = NULL, use = c("pairwise.complete.obs", "complete.obs"))

w_cor(x, weight = NULL, use = c("pairwise.complete.obs", "complete.obs"))

w_pearson(x, weight = NULL, use = c("pairwise.complete.obs",
  "complete.obs"))

w_spearman(x, weight = NULL, use = c("pairwise.complete.obs",
  "complete.obs"))
```

Arguments

x	a numeric vector (matrix/data.frame for correlations) containing the values whose weighted statistics is to be computed.
weight	a vector of weights to use for each element of x. Cases with missing, zero or negative weights will be removed before calculations. If `weight` is missing then unweighted statistics will be computed.
na.rm	a logical value indicating whether NA values should be stripped before the computation proceeds. Note that contrary to base R statistic functions the default value is TRUE (remove missing values).
use	`"pairwise.complete.obs"` (default) or `"complete.obs"`. In the first case the correlation or covariance between each pair of variables is computed using all

complete pairs of observations on those variables. If use is "complete.obs" then missing values are handled by casewise deletion.

Details

If argument of correlation functions is data.frame with variable labels then variables names will be replaced with labels. If this is undesirable behavior use unvr function: w_cor(unvr(x)). Weighted spearman correlation coefficients are calculated with rounded to nearest integer weights. It gives the same result as in SPSS Statistics software. By now this algorithm is not memory efficient.

Value

a numeric value of length one/correlation matrix

Examples

```
data(mtcars)
dfs = mtcars %>% keep(mpg, disp, hp, wt)

with(dfs, w_mean(hp, weight = 1/wt))

# apply labels
dfs = modify(dfs, {
         var_lab(mpg) = "Miles/(US) gallon"
         var_lab(disp) = "Displacement (cu.in.)"
         var_lab(hp) = "Gross horsepower"
         var_lab(wt) = "Weight (1000 lbs)"
})

# weighted correlations with labels
w_cor(dfs, weight = 1/dfs$wt)

# without labels
w_cor(unvr(dfs), weight = 1/dfs$wt)
```

Index