

Package ‘dplyr’

October 11, 2014

Type Package

Version 0.3.0.2

Title A Grammar of Data Manipulation

Description

A fast, consistent tool for working with data frame like objects, both in memory and out of memory.

URL <https://github.com/hadley/dplyr>

BugReports <https://github.com/hadley/dplyr/issues>

Depends R (>= 3.1)

Imports assertthat, utils, R6, Rcpp, magrittr, lazyeval (>= 0.1.8), DBI (>= 0.3)

Suggests RSQLite, RSQLite.extfuns, RMySQL, RPostgreSQL, data.table, testthat, knitr, microbenchmark, ggplot2, mgcv, Lahman (>= 3.0-1), nycflights13, methods

VignetteBuilder knitr

LazyData yes

LinkingTo Rcpp (>= 0.11.3), BH (>= 1.51.0-2)

License MIT + file LICENSE

Collate 'RcppExports.R' 'all-equal.r' 'bench-compare.r' 'cbind.r'
'chain.r' 'cluster.R' 'colwise.R' 'compute-collect.r'
'copy-to.r' 'data-lahman.r' 'data-nasa.r' 'data-nycflights13.r'
'data-temp.r' 'data.r' 'dataframe.R' 'dbi-s3.r' 'desc.r'
'distinct.R' 'do.r' 'dplyr.r' 'explain.r' 'failwith.r' 'funs.R'
'glimpse.R' 'group-by.r' 'group-size.r' 'grouped-df.r'
'grouped-dt.r' 'id.r' 'inline.r' 'join-df.r' 'join-dt.r'
'join-sql.r' 'join.r' 'lead-lag.R' 'location.R' 'manip-cube.r'
'manip-df.r' 'manip-dt.r' 'manip-sql.r' 'manip.r' 'nth-value.R'
'order-by.R' 'over.R' 'partial-eval.r' 'progress.R' 'query.r'
'rank.R' 'rbind.r' 'rowwise.r' 'sample.R' 'select-utils.R'

'select-vars.R' 'sets.r' 'sql-escape.r' 'sql-star.r' 'src-local.r' 'src-mysql.r' 'src-postgres.r' 'src-sql.r'
 'src-sqlite.r' 'src.r' 'tally.R' 'tbl-cube.r' 'tbl-data-frame.R' 'tbl-df.r' 'tbl-dt.r' 'tbl-sql.r' 'tbl.r'
 'top-n.R' 'translate-sql-helpers.r' 'translate-sql-base.r'
 'translate-sql-window.r' 'translate-sql.r' 'type-sum.r'
 'utils-dt.R' 'utils-format.r' 'utils.r' 'view.r' 'zzz.r'

Author Hadley Wickham [aut, cre], Romain Francois [aut], RStudio [cph]

Maintainer Hadley Wickham <hadley@rstudio.com>

NeedsCompilation yes

Repository CRAN

Date/Publication 2014-10-11 07:43:41

R topics documented:

all.equal.tbl_df	3
arrange	4
as.tbl_cube	5
bench_compare	6
between	7
build_sql	8
chain	9
compute	10
copy_to	11
copy_to.src_sql	12
cumall	13
data_frame	13
desc	14
distinct	15
do	16
dplyr	17
explain	18
failwith	19
filter	20
funs	21
glimpse	21
grouped_dt	22
groups	23
group_by	23
group_size	25
join	25
join.tbl_df	27
join.tbl_dt	28
join.tbl_sql	29
lead-lag	31
location	32

mutate	33
n	34
nasa	34
nth	35
n_distinct	36
order_by	36
ranking	37
rbind_list	38
rowwise	39
sample	39
select	40
setops	42
slice	43
src_mysql	44
src_postgres	47
src_sqlite	49
src_tbls	52
summarise	53
summarise_each	54
tally	55
tbl	56
tbl_cube	56
tbl_df	58
tbl_dt	59
tbl_vars	61
top_n	61
translate_sql	62
Index	65

all.equal.tbl_df	<i>Provide a useful implementation of all.equal for data.frames.</i>
------------------	--

Description

Provide a useful implementation of all.equal for data.frames.

Usage

```
## S3 method for class 'tbl_df'
all.equal(target, current, ignore_col_order = TRUE,
  ignore_row_order = TRUE, convert = FALSE, ...)

## S3 method for class 'tbl_dt'
all.equal(target, current, ignore_col_order = TRUE,
  ignore_row_order = TRUE, convert = FALSE, ...)
```

Arguments

target,current two data frames to compare
 ignore_col_order should order of columns be ignored?
 ignore_row_order should order of rows be ignored?
 convert Should similar classes be converted? Currently this will convert factor to character and integer to double.
 ... Ignored. Needed for compatibility with the generic.

Value

TRUE if equal, otherwise a character vector describing the first reason why they're not equal. Use [isTRUE](#) if using the result in an if expression.

Examples

```
scramble <- function(x) x[sample(nrow(x)), sample(ncol(x))]  
  
# By default, ordering of rows and columns ignored  
mtcars_df <- tbl_df(mtcars)  
all.equal(mtcars_df, scramble(mtcars_df))  
  
# But those can be overridden if desired  
all.equal(mtcars_df, scramble(mtcars_df), ignore_col_order = FALSE)  
all.equal(mtcars_df, scramble(mtcars_df), ignore_row_order = FALSE)
```

arrange	<i>Arrange rows by variables.</i>
---------	-----------------------------------

Description

Use [desc](#) to sort a variable in descending order.

Usage

```
arrange(.data, ...)  
  
arrange_(.data, ..., .dots)
```

Arguments

.data A tbl. All main verbs are S3 generics and provide methods for [tbl_df](#), [tbl_dt](#) and [tbl_sql](#).
 ... Comma separated list of unquoted variable names. Use [desc](#) to sort a variable in descending order.
 .dots Used to work around non-standard evaluation. See [vignette\("nse"\)](#) for details.

Value

An object of the same class as `.data`.

Data frame row names are silently dropped. To preserve, convert to an explicit variable.

Locales

Note that for local data frames, the ordering is done in C++ code which does not have access to the local specific ordering usually done in R. This means that strings are ordered as if in the C locale.

See Also

Other `single.table.verbs`: [filter](#), [filter_](#); [mutate](#), [mutate_](#), [transmute](#), [transmute_](#); [rename](#), [rename_](#), [select](#), [select_](#); [slice](#), [slice_](#); [summarise](#), [summarise_](#), [summarize](#), [summarize_](#)

Examples

```
arrange(mtcars, cyl, disp)
arrange(mtcars, desc(disp))
```

as.tbl_cube

Coerce an existing data structure into a tbl_cube

Description

Coerce an existing data structure into a `tbl_cube`

Usage

```
as.tbl_cube(x, ...)

## S3 method for class 'array'
as.tbl_cube(x, met_name = deparse(substitute(x)),
  dim_names = names(dimnames(x)), ...)

## S3 method for class 'table'
as.tbl_cube(x, met_name = deparse(substitute(x)),
  dim_names = names(dimnames(x)), ...)

## S3 method for class 'matrix'
as.tbl_cube(x, met_name = deparse(substitute(x)),
  dim_names = names(dimnames(x)), ...)

## S3 method for class 'data.frame'
as.tbl_cube(x, dim_names, ...)
```

Arguments

x	an object to convert. Built in methods will convert arrays, tables and data frames.
...	Passed on to individual methods; otherwise ignored.
met_name	a string to use as the name for the metric
dim_names	names of the dimesions. Defaults to the names of the dimnames .

bench_compare	<i>Evaluate, compare, benchmark operations of a set of srcs.</i>
---------------	--

Description

These functions support the comparison of results and timings across multiple sources.

Usage

```
bench_tbls(tbls, op, ..., times = 10)

compare_tbls(tbls, op, ref = NULL, compare = equal_data_frame, ...)

eval_tbls(tbls, op)
```

Arguments

tbls	A list of tbls .
op	A function with a single argument, called often with each element of tbls.
times	For benchmarking, the number of times each operation is repeated.
ref	For checking, an data frame to test results against. If not supplied, defaults to the results from the first src.
compare	A function used to compare the results. Defaults to <code>equal_data_frame</code> which ignores the order of rows and columns.
...	For <code>compare_tbls</code> : additional parameters passed on the compare function For <code>bench_tbls</code> : additional benchmarks to run.

Value

`eval_tbls`: a list of data frames.

`compare_tbls`: an invisible TRUE on success, otherwise an error is thrown.

`bench_tbls`: an object of class [microbenchmark](#)

See Also

[src_local](#) for working with local data

Examples

```
if (require("microbenchmark") && has_lahman()) {
  lahman_local <- lahman_srcs("df", "dt")
  teams <- lapply(lahman_local, function(x) x %>% tbl("Teams"))

  compare_tbls(teams, function(x) x %>% filter(yearID == 2010))
  bench_tbls(teams, function(x) x %>% filter(yearID == 2010))

  # You can also supply arbitrary additional arguments to bench_tbls
  # if there are other operations you'd like to compare.
  bench_tbls(teams, function(x) x %>% filter(yearID == 2010),
    base = subset(Lahman::Teams, yearID == 2010))

  # A more complicated example using multiple tables
  setup <- function(src) {
    list(
      src %>% tbl("Batting") %>% filter(stint == 1) %>% select(playerID:H),
      src %>% tbl("Master") %>% select(playerID, birthYear)
    )
  }
  two_tables <- lapply(lahman_local, setup)

  op <- function(tbls) {
    semi_join(tbls[[1]], tbls[[2]], by = "playerID")
  }
  # compare_tbls(two_tables, op)
  bench_tbls(two_tables, op, times = 2)

}
```

between

Do values in a numeric vector fail in specified range?

Description

This is a shortcut for `x >= left & x <= right`, implemented efficiently in C++ for local values, and translated to the appropriate SQL for remote tables.

Usage

```
between(x, left, right)
```

Arguments

<code>x</code>	A numeric vector of values
<code>left, right</code>	Boundary values

Examples

```
x <- rnorm(1e2)
x[between(x, -1, 1)]
```

build_sql	<i>Build a SQL string.</i>
-----------	----------------------------

Description

This is a convenience function that should prevent sql injection attacks (which in the context of dplyr are most likely to be accidental not deliberate) by automatically escaping all expressions in the input, while treating bare strings as sql. This is unlikely to prevent any serious attack, but should make it unlikely that you produce invalid sql.

Usage

```
build_sql(..., .env = parent.frame(), con = NULL)
```

Arguments

...	input to convert to SQL. Use <code>sql</code> to preserve user input as is (dangerous), and <code>ident</code> to label user input as sql identifiers (safe)
.env	the environment in which to evaluate the arguments. Should not be needed in typical use.
con	database connection; used to select correct quoting characters.

Examples

```
build_sql("SELECT * FROM TABLE")
x <- "TABLE"
build_sql("SELECT * FROM ", x)
build_sql("SELECT * FROM ", ident(x))
build_sql("SELECT * FROM ", sql(x))

# http://xkcd.com/327/
name <- "Robert'); DROP TABLE Students;--"
build_sql("INSERT INTO Students (Name) VALUES (", name, ")")
```

chain	<i>Chain together multiple operations.</i>
-------	--

Description

The downside of the functional nature of dplyr is that when you combine multiple data manipulation operations, you have to read from the inside out and the arguments may be very distant to the function call. These functions providing an alternative way of calling dplyr (and other data manipulation) functions that you read can from left to right.

Usage

```
chain(..., env = parent.frame())

chain_q(calls, env = parent.frame())

lhs %.% rhs

lhs %>% rhs
```

Arguments

<code>..., calls</code>	A sequence of data transformations, starting with a dataset. The first argument of each call should be omitted - the value of the previous step will be substituted in automatically. Use <code>chain</code> and <code>...</code> when working interactive; use <code>chain_q</code> and <code>calls</code> when calling from another function.
<code>env</code>	Environment in which to evaluation expressions. In ordinary operation you should not need to set this parameter.
<code>lhs, rhs</code>	A dataset and function to apply to it

Details

The functions work via simple substitution so that `x %.% f(y)` is translated into `f(x, y)`.

Deprecation

`chain` was deprecated in version 0.2, and will be removed in 0.3. It was removed in the interest of making dplyr code more standardised and `%.%` is much more popular.

Examples

```
# If you're performing many operations you can either do step by step
if (require("nycflights13")) {
  a1 <- group_by(flights, year, month, day)
  a2 <- select(a1, arr_delay, dep_delay)
  a3 <- summarise(a2,
    arr = mean(arr_delay, na.rm = TRUE),
```

```

    dep = mean(dep_delay, na.rm = TRUE))
a4 <- filter(a3, arr > 30 | dep > 30)

# If you don't want to save the intermediate results, you need to
# wrap the functions:
filter(
  summarise(
    select(
      group_by(flights, year, month, day),
      arr_delay, dep_delay
    ),
    arr = mean(arr_delay, na.rm = TRUE),
    dep = mean(dep_delay, na.rm = TRUE)
  ),
  arr > 30 | dep > 30
)

# This is difficult to read because the order of the operations is from
# inside to out, and the arguments are a long way away from the function.
# Alternatively you can use chain or %>% to sequence the operations
# linearly:

flights %>%
  group_by(year, month, day) %>%
  select(arr_delay, dep_delay) %>%
  summarise(
    arr = mean(arr_delay, na.rm = TRUE),
    dep = mean(dep_delay, na.rm = TRUE)
  ) %>%
  filter(arr > 30 | dep > 30)
}

```

compute

Compute a lazy tbl.

Description

compute forces computation of lazy tbls, leaving data in the remote source. collect also forces computation, but will bring data back into an R data.frame (stored in a [tbl_df](#)). collapse doesn't force computation, but collapses a complex tbl into a form that additional restrictions can be placed on.

Usage

```
compute(x, name = random_table_name(), ...)
```

```
collect(x, ...)
```

```
collapse(x, ...)
```

```
## S3 method for class 'tbl_sql'
compute(x, name = random_table_name(), temporary = TRUE,
  ...)
```

Arguments

x	a data tbl
name	name of temporary table on database.
...	other arguments passed on to methods
temporary	if TRUE, will create a temporary table that is local to this connection and will be automatically deleted when the connection expires

Grouping

compute and collect preserve grouping, collapse drops it.

See Also

[copy_to](#) which is the conceptual opposite: it takes a local data frame and makes it available to the remote source.

Examples

```
if (require("RSQLite") && has_lahman("sqlite")) {
  batting <- tbl(lahman_sqlite(), "Batting")
  remote <- select(filter(batting, yearID > 2010 && stint == 1), playerID:H)
  remote2 <- collapse(remote)
  cached <- compute(remote)
  local <- collect(remote)
}
```

copy_to	<i>Copy a local data frame to a remote src.</i>
---------	---

Description

This uploads a local data frame into a remote data source, creating the table definition as needed. Wherever possible, the new object will be temporary, limited to the current connection to the source.

Usage

```
copy_to(dest, df, name = deparse(substitute(df)), ...)
```

Arguments

dest	remote data source
df	local data frame
name	name for new remote table.
...	other parameters passed to methods.

Value

a tbl object in the remote source

copy_to.src_sql	<i>Copy a local data fram to a sqlite src.</i>
-----------------	--

Description

This standard method works for all sql sources.

Usage

```
## S3 method for class 'src_sql'
copy_to(dest, df, name = deparse(substitute(df)),
  types = NULL, temporary = TRUE, indexes = NULL, analyze = TRUE, ...)
```

Arguments

dest	remote data source
df	local data frame
name	name for new remote table.
types	a character vector giving variable types to use for the columns. See http://www.sqlite.org/datatype3.html for available types.
temporary	if TRUE, will create a temporary table that is local to this connection and will be automatically deleted when the connection expires
indexes	a list of character vectors. Each element of the list will create a new index.
analyze	if TRUE (the default), will automatically ANALYZE the new table so that the query optimiser has useful information.
...	other parameters passed to methods.

Value

a sqlite `tbl` object

Examples

```

if (require("RSQLite") && require("RSQLite.extfuns")) {
  db <- src_sqlite(tempfile(), create = TRUE)

  iris2 <- copy_to(db, iris)
  mtcars$model <- rownames(mtcars)
  mtcars2 <- copy_to(db, mtcars, indexes = list("model"))

  explain(filter(mtcars2, model == "Hornet 4 Drive"))

  # Note that tables are temporary by default, so they're not
  # visible from other connections to the same database.
  src_tbls(db)
  db2 <- src_sqlite(db$path)
  src_tbls(db2)
}

```

cumall

*Cumulative versions of any, all, and mean***Description**

dplyr adds cumall, cumany, and cummean to complete R's set of cumulate functions to match the aggregation functions available in most databases

Usage

cumall(x)

cumany(x)

cummean(x)

Arguments

x For cumall & cumany, a logical vector; for cummean an integer or numeric vector

data_frame

Build a data frame.

Description

A trimmed down version of `data.frame` that:

1. Never coerces inputs (i.e. strings stay as strings!).
2. Never adds `row.names`.
3. Never munges column names.
4. Only recycles length 1 inputs.
5. Evaluates its arguments lazily and in order.
6. Adds `tbl_df` class to output.

Usage

```
data_frame(...)
```

```
data_frame_(columns)
```

Arguments

... A set of named arguments

columns A [lazy_dots](#).

Examples

```
a <- 1:5
data_frame(a, b = a * 2)
data_frame(a, b = a * 2, c = 1)
data_frame(x = runif(10), y = x * 2)

# data_frame never coerces its inputs
str(data_frame(letters))
str(data_frame(x = diag(5)))

# or munges column names
data_frame(`a + b` = 1:5)
```

desc

Descending order.

Description

Transform a vector into a format that will be sorted in descending order.

Usage

```
desc(x)
```

Arguments

x vector to transform

Examples

```
desc(1:10)
desc(factor(letters))
first_day <- seq(as.Date("1910/1/1"), as.Date("1920/1/1"), "years")
desc(first_day)
```

distinct	<i>Select distinct/unique rows.</i>
----------	-------------------------------------

Description

Retain only unique/distinct rows from an input tbl. This is an efficient version of [unique](#). `distinct()` is best-suited for interactive use, `distinct_()` for calling from a function.

Usage

```
distinct(.data, ...)

distinct_(.data, ..., .dots)
```

Arguments

.data	a tbl
...	Variables to use when determining uniqueness. If there are multiple rows for a given combination of inputs, only the first row will be preserved.
.dots	Used to work around non-standard evaluation. See vignette("nse") for details.

Examples

```
df <- data.frame(
  x = sample(10, 100, rep = TRUE),
  y = sample(10, 100, rep = TRUE)
)
nrow(df)
nrow(distinct(df))
distinct(df, x)
distinct(df, y)

# You can also use distinct on computed variables
distinct(df, diff = abs(x - y))
```

do	<i>Do arbitrary operations on a tbl.</i>
----	--

Description

This is a general purpose complement to the specialised manipulation functions [filter](#), [select](#), [mutate](#), [summarise](#) and [arrange](#). You can use `do` to perform arbitrary computation, returning either a data frame or arbitrary objects which will be stored in a list. This is particularly useful when working with models: you can fit models per group with `do` and then flexibly extract components with either another `do` or `summarise`.

Usage

```
do(.data, ...)
```

```
## S3 method for class 'tbl_sql'
```

```
do(.data, ..., .chunk_size = 10000L)
```

Arguments

<code>.data</code>	a <code>tbl</code>
<code>...</code>	Expressions to apply to each group. If named, results will be stored in a new column. If unnamed, should return a data frame. You can use <code>.</code> to refer to the current group. You can not mix named and unnamed arguments.
<code>.chunk_size</code>	The size of each chunk to pull into R. If this number is too big, the process will be slow because R has to allocate and free a lot of memory. If it's too small, it will be slow, because of the overhead of talking to the database.

Value

`do` always returns a data frame. The first columns in the data frame will be the labels, the others will be computed from `...`. Named arguments become list-columns, with one element for each group; unnamed elements must be data frames and labels will be duplicated accordingly.

Groups are preserved for a single unnamed input. This is different to [summarise](#) because `do` generally does not reduce the complexity of the data, it just expresses it in a special way. For multiple named inputs, the output is grouped by row with [rowwise](#). This allows other verbs to work in an intuitive way.

Connection to plyr

If you're familiar with `plyr`, `do` with named arguments is basically equivalent to `dply`, and `do` with a single unnamed argument is basically equivalent to `ldply`. However, instead of storing labels in a separate attribute, the result is always a data frame. This means that `summarise` applied to the result of `do` can act like `ldply`.

Examples

```

by_cyl <- group_by(mtcars, cyl)
do(by_cyl, head(., 2))

models <- by_cyl %>% do(mod = lm(mpg ~ disp, data = .))
models

summarise(models, rsq = summary(mod)$r.squared)
models %>% do(data.frame(coef = coef(.$mod)))
models %>% do(data.frame(
  var = names(coef(.$mod)),
  coef(summary(.$mod)))
)

models <- by_cyl %>% do(
  mod_linear = lm(mpg ~ disp, data = .),
  mod_quad = lm(mpg ~ poly(disp, 2), data = .)
)
models
compare <- models %>% do(aov = anova(.$mod_linear, .$mod_quad))
# compare %>% summarise(p.value = aov$`Pr(>F)`)

if (require("nycflights13")) {
# You can use it to do any arbitrary computation, like fitting a linear
# model. Let's explore how carrier departure delays vary over the time
carriers <- group_by(flights, carrier)
group_size(carriers)

mods <- do(carriers, mod = lm(arr_delay ~ dep_time, data = .))
mods %>% do(as.data.frame(coef(.$mod)))
mods %>% summarise(rsq = summary(mod)$r.squared)

## Not run:
# This longer example shows the progress bar in action
by_dest <- flights %>% group_by(dest) %>% filter(n() > 100)
library(mgcv)
by_dest %>% do(smooth = gam(arr_delay ~ s(dep_time) + month, data = .))

## End(Not run)
}

```

Description

The dplyr package.

explain	<i>Explain details of an tbl.</i>
---------	-----------------------------------

Description

This is a generic function which gives more details about an object than `print`, and is more focussed on human readable output than `str`.

Usage

```
explain(x, ...)
```

```
show_query(x)
```

Arguments

<code>x</code>	An object to explain
<code>...</code>	Other parameters possibly used by generic

Databases

Explaining a `tbl_sql` will run the SQL EXPLAIN command which will describe the query plan. This requires a little bit of knowledge about how EXPLAIN works for your database, but is very useful for diagnosing performance problems.

Examples

```
if (require("RSQLite") && has_lahman("sqlite")) {

  batting <- tbl(lahman_sqlite(), "Batting")
  batting %>% show_query()
  batting %>% explain()

  # The batting database has indices on all ID variables:
  # SQLite automatically picks the most restrictive index
  batting %>% filter(lgID == "NL" & yearID == 2000L) %>% explain()

  # OR's will use multiple indexes
  batting %>% filter(lgID == "NL" | yearID == 2000) %>% explain()

  # Joins will use indexes in both tables
  teams <- tbl(lahman_sqlite(), "Teams")
  batting %>% left_join(teams, c("yearID", "teamID")) %>% explain()
}
```

failwith	<i>Fail with specified value.</i>
----------	-----------------------------------

Description

Modify a function so that it returns a default value when there is an error.

Usage

```
failwith(default = NULL, f, quiet = FALSE)
```

Arguments

default	default value
f	function
quiet	all error messages be suppressed?

Value

a function

See Also

[try_default](#)

Examples

```
f <- function(x) if (x == 1) stop("Error!") else 1
## Not run:
f(1)
f(2)

## End(Not run)

safef <- failwith(NULL, f)
safef(1)
safef(2)
```

filter	<i>Return rows with matching conditions.</i>
--------	--

Description

Return rows with matching conditions.

Usage

```
filter(.data, ...)
```

```
filter_(.data, ..., .dots)
```

Arguments

<code>.data</code>	A tbl. All main verbs are S3 generics and provide methods for tbl_df , tbl_dt and tbl_sql .
<code>...</code>	Logical predicates. Multiple conditions are combined with &.
<code>.dots</code>	Used to work around non-standard evaluation. See vignette("nse") for details.

Value

An object of the same class as `.data`.

Data frame row names are silently dropped. To preserve, convert to an explicit variable.

See Also

Other `single.table.verbs`: [arrange](#), [arrange_](#); [mutate](#), [mutate_](#), [transmute](#), [transmute_](#); [rename](#), [rename_](#), [select](#), [select_](#); [slice](#), [slice_](#); [summarise](#), [summarise_](#), [summarize](#), [summarize_](#)

Examples

```
filter(mtcars, cyl == 8)
filter(mtcars, cyl < 6)
```

funs	Create a list of functions calls.
------	-----------------------------------

Description

funs provides a flexible to generate a named list of functions for input to other functions like colwise.

Usage

```
funs(...)  
  
funs_(dots)
```

Arguments

dots, ... A list of functions specified by:

- Their name, "mean"
- The function itself, mean
- A call to the function with . as a dummy parameter, mean(., na.rm = TRUE)

Examples

```
funs(mean, "mean", mean(., na.rm = TRUE))  
  
# Override default names  
funs(m1 = mean, m2 = "mean", m3 = mean(., na.rm = TRUE))  
  
# If you have a function names in a vector, use funs_q  
fs <- c("min", "max")  
funs_(fs)
```

glimpse	Get a glimpse of your data.
---------	-----------------------------

Description

This is like a transposed version of print: columns run down the page, and data runs across. This makes it possible to see every column in a data frame. It's a little like [str](#) applied to a data frame but it tries to show you as much data as possible. (And it always shows the underlying data, even when applied to a remote data source.)

Usage

```
glimpse(tbl, width = getOption("width"))
```

Arguments

tbl	A data table
width	Width of output: defaults to the width of the console.

Examples

```
glimpse(mtcars)

if (require("RSQLite") && has_lahman("sqlite")) {
  batting <- tbl(lahman_sqlite(), "Batting")
  glimpse(batting)
}
```

grouped_dt	<i>A grouped data table.</i>
------------	------------------------------

Description

The easiest way to create a grouped data table is to call the `group_by` method on a data table or `tbl`: this will take care of capturing the unevaluated expressions for you.

Usage

```
grouped_dt(data, vars, copy = TRUE)

is.grouped_dt(x)
```

Arguments

data	a tbl or data frame.
vars	a list of quoted variables.
copy	If TRUE, will make copy of input.
x	an object to check

Examples

```
if (require("data.table") && require("nycflights13")) {
  flights_dt <- tbl_dt(flights)
  group_size(group_by(flights_dt, year, month, day))
  group_size(group_by(flights_dt, dest))

  monthly <- group_by(flights_dt, month)
  summarise(monthly, n = n(), delay = mean(arr_delay))
}
```

groups	<i>Get/set the grouping variables for tbl.</i>
--------	--

Description

These functions do not perform non-standard evaluation, and so are useful when programming against tbl objects. ungroup is a convenient inline way of removing existing grouping.

Usage

```
groups(x)
```

```
ungroup(x)
```

Arguments

x	data tbl
---	--------------------------

Examples

```
grouped <- group_by(mtcars, cyl)
groups(grouped)
groups(ungroup(grouped))
```

group_by	<i>Group a tbl by one or more variables.</i>
----------	--

Description

Most data operations are useful done on groups defined by variables in the the dataset. The group_by function takes an existing tbl and converts it into a grouped tbl where operations are performed "by group".

Usage

```
group_by(.data, ..., add = FALSE)
```

```
group_by_(.data, ..., .dots, add = FALSE)
```

Arguments

.data	a tbl
...	variables to group by. All tbls accept variable names, some will also accept functions of variables. Duplicated groups will be silently dropped.
add	By default, when add = FALSE, group_by will override existing groups. To instead add to the existing groups, use add = TRUE
.dots	Used to work around non-standard evaluation. See vignette("nse") for details.

Tbl types

group_by is an S3 generic with methods for the three built-in tbls. See the help for the corresponding classes and their manip methods for more details:

- data.frame: [grouped_df](#)
- data.table: [grouped_dt](#)
- SQLite: [src_sqlite](#)
- PostgreSQL: [src_postgres](#)
- MySQL: [src_mysql](#)

See Also

[ungroup](#) for the inverse operation, [groups](#) for accessors that don't do special evaluation.

Examples

```
by_cyl <- group_by(mtcars, cyl)
summarise(by_cyl, mean(displ), mean(hp))
filter(by_cyl, displ == max(displ))

# summarise peels off a single layer of grouping
by_vs_am <- group_by(mtcars, vs, am)
by_vs <- summarise(by_vs_am, n = n())
by_vs
summarise(by_vs, n = sum(n))
# use ungroup() to remove if not wanted
summarise(ungroup(by_vs), n = sum(n))

# You can group by expressions: this is just short-hand for
# a mutate/rename followed by a simple group_by
group_by(mtcars, vsam = vs + am)
group_by(mtcars, vs2 = vs)

# You can also group by a constant, but it's not very useful
group_by(mtcars, "vs")

# By default, group_by sets groups. Use add = TRUE to add groups
groups(group_by(by_cyl, vs, am))
groups(group_by(by_cyl, vs, am, add = TRUE))

# Duplicate groups are silently dropped
groups(group_by(by_cyl, cyl, cyl))
```

group_size	<i>Calculate group sizes.</i>
------------	-------------------------------

Description

Calculate group sizes.

Usage

```
group_size(x)
```

```
n_groups(x)
```

Arguments

x a grouped tbl

Examples

```
if (require("nycflights13")) {  
  
  by_day <- flights %>% group_by(year, month, day)  
  n_groups(by_day)  
  group_size(by_day)  
  
  by_dest <- flights %>% group_by(dest)  
  n_groups(by_dest)  
  group_size(by_dest)  
}
```

join	<i>Join two tbls together.</i>
------	--------------------------------

Description

These are generic functions that dispatch to individual tbl methods - see the method documentation for details of individual data sources. x and y should usually be from the same data source, but if copy is TRUE, y will automatically be copied to the same source as x - this may be an expensive operation.

Usage

```
inner_join(x, y, by = NULL, copy = FALSE, ...)
```

```
left_join(x, y, by = NULL, copy = FALSE, ...)
```

```
semi_join(x, y, by = NULL, copy = FALSE, ...)
```

```
anti_join(x, y, by = NULL, copy = FALSE, ...)
```

Arguments

<code>x, y</code>	tbls to join
<code>by</code>	a character vector of variables to join by. If <code>NULL</code> , the default, join will do a natural join, using all variables with common names across the two tables. A message lists the variables so that you can check they're right. To join by different variables on <code>x</code> and <code>y</code> use a named vector. For example, <code>by = c("a" = "b")</code> will match <code>x.a</code> to <code>y.b</code> .
<code>copy</code>	If <code>x</code> and <code>y</code> are not from the same data source, and <code>copy</code> is <code>TRUE</code> , then <code>y</code> will be copied into the same src as <code>x</code> . This allows you to join tables across srcs, but it is a potentially expensive operation so you must opt into it.
<code>...</code>	other parameters passed onto methods

Join types

Currently dplyr supports four join types:

`inner_join` return all rows from `x` where there are matching values in `y`, and all columns from `x` and `y`. If there are multiple matches between `x` and `y`, all combination of the matches are returned.

`left_join` return all rows from `x`, and all columns from `x` and `y`. If there are multiple matches between `x` and `y`, all combination of the matches are returned.

`semi_join` return all rows from `x` where there are matching values in `y`, keeping just columns from `x`.

A semi join differs from an inner join because an inner join will return one row of `x` for each matching row of `y`, where a semi join will never duplicate rows of `x`.

`anti_join` return all rows from `x` where there are not matching values in `y`, keeping just columns from `x`

Grouping

Groups are ignored for the purpose of joining, but the result preserves the grouping of `x`.

join.tbl_df	<i>Join data frame tbls.</i>
-------------	------------------------------

Description

See [join](#) for a description of the general purpose of the functions.

Usage

```
## S3 method for class 'tbl_df'
inner_join(x, y, by = NULL, copy = FALSE, ...)

## S3 method for class 'tbl_df'
left_join(x, y, by = NULL, copy = FALSE, ...)

## S3 method for class 'tbl_df'
semi_join(x, y, by = NULL, copy = FALSE, ...)

## S3 method for class 'tbl_df'
anti_join(x, y, by = NULL, copy = FALSE, ...)
```

Arguments

x,y	tbls to join
by	a character vector of variables to join by. If NULL, the default, join will do a natural join, using all variables with common names across the two tables. A message lists the variables so that you can check they're right - to suppress the message, supply a character vector.
copy	If y is not a data frame or tbl_df and copy is TRUE, y will be converted into a data frame
...	included for compatibility with the generic; otherwise ignored.

Examples

```
if (require("Lahman")) {
  batting_df <- tbl_df(Batting)
  person_df <- tbl_df(Master)

  uperson_df <- tbl_df(Master[!duplicated(Master$playerID), ])

  # Inner join: match batting and person data
  inner_join(batting_df, person_df)
  inner_join(batting_df, uperson_df)

  # Left join: match, but preserve batting data
  left_join(batting_df, uperson_df)
```

```
# Anti join: find batters without person data
anti_join(batting_df, person_df)
# or people who didn't bat
anti_join(person_df, batting_df)
}
```

join.tbl_dt

Join data table tbls.

Description

See [join](#) for a description of the general purpose of the functions.

Usage

```
## S3 method for class 'data.table'
inner_join(x, y, by = NULL, copy = FALSE, ...)

## S3 method for class 'data.table'
left_join(x, y, by = NULL, copy = FALSE, ...)

## S3 method for class 'data.table'
semi_join(x, y, by = NULL, copy = FALSE, ...)

## S3 method for class 'data.table'
anti_join(x, y, by = NULL, copy = FALSE, ...)
```

Arguments

x,y	tbls to join
by	a character vector of variables to join by. If NULL, the default, join will do a natural join, using all variables with common names across the two tables. A message lists the variables so that you can check they're right. To join by different variables on x and y use a named vector. For example, <code>by = c("a" = "b")</code> will match <code>x.a</code> to <code>y.b</code> .
copy	If x and y are not from the same data source, and copy is TRUE, then y will be copied into the same src as x. This allows you to join tables across srcs, but it is a potentially expensive operation so you must opt into it.
...	Included for compatibility with generic; otherwise ignored.

Examples

```
if (require("data.table") && require("Lahman")) {
  batting_dt <- tbl_dt(Batting)
  person_dt <- tbl_dt(Master)

  # Inner join: match batting and person data
```

```

inner_join(batting_dt, person_dt)

# Left join: keep batting data even if person missing
left_join(batting_dt, person_dt)

# Semi-join: find batting data for top 4 teams, 2010:2012
grid <- expand.grid(
  teamID = c("WAS", "ATL", "PHI", "NYA"),
  yearID = 2010:2012)
top4 <- semi_join(batting_dt, grid, copy = TRUE)

# Anti-join: find batting data with out player data
anti_join(batting_dt, person_dt)
}

```

join.tbl_sql

*Join sql tbls.***Description**

See [join](#) for a description of the general purpose of the functions.

Usage

```

## S3 method for class 'tbl_sql'
inner_join(x, y, by = NULL, copy = FALSE,
  auto_index = FALSE, ...)

## S3 method for class 'tbl_sql'
left_join(x, y, by = NULL, copy = FALSE,
  auto_index = FALSE, ...)

## S3 method for class 'tbl_sql'
semi_join(x, y, by = NULL, copy = FALSE,
  auto_index = FALSE, ...)

## S3 method for class 'tbl_sql'
anti_join(x, y, by = NULL, copy = FALSE,
  auto_index = FALSE, ...)

```

Arguments

<code>x,y</code>	tbls to join
<code>by</code>	<p>a character vector of variables to join by. If NULL, the default, join will do a natural join, using all variables with common names across the two tables. A message lists the variables so that you can check they're right.</p> <p>To join by different variables on x and y use a named vector. For example, <code>by = c("a" = "b")</code> will match x.a to y.b.</p>

copy	If x and y are not from the same data source, and copy is TRUE, then y will be copied into a temporary table in same database as x. join will automatically run ANALYZE on the created table in the hope that this will make you queries as efficient as possible by giving more data to the query planner. This allows you to join tables across srcs, but it's potentially expensive operation so you must opt into it.
auto_index	if copy is TRUE, automatically create indices for the variables in by. This may speed up the join if there are matching indexes in x.
...	other parameters passed onto methods

Implementation notes

Semi-joins are implemented using WHERE EXISTS, and anti-joins with WHERE NOT EXISTS. Support for semi-joins is somewhat partial: you can only create semi joins where the x and y columns are compared with = not with more general operators.

Examples

```
if (require("RSQLite") && has_lahman("sqlite")) {

# Left joins -----
batting <- tbl(lahman_sqlite(), "Batting")
team_info <- select(tbl(lahman_sqlite(), "Teams"), yearID, lgID, teamID, G, R:H)

# Combine player and whole team statistics
first_stint <- select(filter(batting, stint == 1), playerID:H)
both <- left_join(first_stint, team_info, type = "inner", by = c("yearID", "teamID", "lgID"))
head(both)
explain(both)

# Join with a local data frame
grid <- expand.grid(
  teamID = c("WAS", "ATL", "PHI", "NYA"),
  yearID = 2010:2012)
top4a <- left_join(batting, grid, copy = TRUE)
explain(top4a)

# Indices don't really help here because there's no matching index on
# batting
top4b <- left_join(batting, grid, copy = TRUE, auto_index = TRUE)
explain(top4b)

# Semi-joins -----

people <- tbl(lahman_sqlite(), "Master")

# All people in hall of fame
hof <- tbl(lahman_sqlite(), "HallOfFame")
semi_join(people, hof)
```

```

# All people not in the hall of fame
anti_join(people, hof)

# Find all managers
manager <- tbl(lahman_sqlite(), "Managers")
semi_join(people, manager)

# Find all managers in hall of fame
famous_manager <- semi_join(semi_join(people, manager), hof)
famous_manager
explain(famous_manager)

# Anti-joins -----

# batters without person covariates
anti_join(batting, people)
}

```

lead-lag

Lead and lag.

Description

Lead and lag are useful for comparing values offset by a constant (e.g. the previous or next value)

Usage

```
lead(x, n = 1L, default = NA, order_by = NULL, ...)
```

```
## Default S3 method:
```

```
lag(x, n = 1L, default = NA, order_by = NULL, ...)
```

Arguments

x	a vector of values
n	a positive integer of length 1, giving the number of positions to lead or lag by
default	value used for non-existent rows. Defaults to NA.
order_by	override the default ordering to use another vector
...	Needed for compatibility with lag generic.

Examples

```

lead(1:10, 1)
lead(1:10, 2)

lag(1:10, 1)
lead(1:10, 1)

```

```

x <- runif(5)
cbind(ahead = lead(x), x, behind = lag(x))

# Use order_by if data not already ordered
df <- data.frame(year = 2000:2005, value = (0:5) ^ 2)
scrambled <- df[sample(nrow(df)), ]

wrong <- mutate(scrambled, prev = lag(value))
arrange(wrong, year)

right <- mutate(scrambled, prev = lag(value, order_by = year))
arrange(right, year)

```

location

Print the location in memory of a data frame

Description

This is useful for understand how and when dplyr makes copies of data frames

Usage

```
location(df)
```

```
changes(x, y)
```

Arguments

```
df,          a data frame
x,y          two data frames to compare
```

Examples

```

location(mtcars)

mtcars2 <- mutate(mtcars, cyl2 = cyl * 2)
location(mtcars2)

changes(mtcars, mtcars)
changes(mtcars, mtcars2)

```

mutate	<i>Add new variables.</i>
--------	---------------------------

Description

Mutate adds new variables and preserves existing; transmute drops existing variables.

Usage

```
mutate(.data, ...)  
  
mutate_(.data, ..., .dots)  
  
transmute(.data, ...)  
  
transmute_(.data, ..., .dots)
```

Arguments

.data	A tbl. All main verbs are S3 generics and provide methods for tbl_df , tbl_dt and tbl_sql .
...	Name-value pairs of expressions. Use NULL to drop a variable.
.dots	Used to work around non-standard evaluation. See vignette("nse") for details.

Value

An object of the same class as .data.

Data frame row names are silently dropped. To preserve, convert to an explicit variable.

See Also

Other single.table.verbs: [arrange](#), [arrange_](#); [filter](#), [filter_](#); [rename](#), [rename_](#), [select](#), [select_](#); [slice](#), [slice_](#); [summarise](#), [summarise_](#), [summarize](#), [summarize_](#)

Examples

```
mutate(mtcars, displ_l = disp / 61.0237)  
transmute(mtcars, displ_l = disp / 61.0237)  
  
mutate(mtcars, cyl = NULL)
```

n	<i>The number of observations in the current group.</i>
---	---

Description

This function is implemented special for each data source and can only be used from within [summarise](#), [mutate](#) and [filter](#)

Usage

```
n()
```

Examples

```
if (require("nycflights13")) {
  carriers <- group_by(flights, carrier)
  summarise(carriers, n())
  mutate(carriers, n = n())
  filter(carriers, n() < 100)
}
```

nasa	<i>NASA spatio-temporal data</i>
------	----------------------------------

Description

This data comes from the ASA 2007 data expo, <http://stat-computing.org/dataexpo/2006/>. The data are geographic and atmospheric measures on a very coarse 24 by 24 grid covering Central America. The variables are: temperature (surface and air), ozone, air pressure, and cloud cover (low, mid, and high). All variables are monthly averages, with observations for Jan 1995 to Dec 2000. These data were obtained from the NASA Langley Research Center Atmospheric Sciences Data Center (with permission; see important copyright terms below).

Usage

```
nasa
```

Format

A [tbl_cube](#) with 41,472 observations.

Dimensions

- lat, long: latitude and longitude
- year, month: month and year

Measures

- cloudlow, cloudmed, cloudhigh: cloud cover at three heights
- ozone
- surftemp and temperature
- pressure

Examples

```
nasa
```

nth	<i>Extract the first, last or nth value from a vector.</i>
-----	--

Description

These are straightforward wrappers around `[[`. The main advantage is that you can provide an optional secondary vector that defines the ordering, and provide a default value to use when the input is shorter than expected.

Usage

```
nth(x, n, order_by = NULL, default = default_missing(x))
```

```
first(x, order_by = NULL, default = default_missing(x))
```

```
last(x, order_by = NULL, default = default_missing(x))
```

Arguments

x	A vector
n	For <code>nth_value</code> , a single integer specifying the position. If a numeric is supplied, it will be silently truncated.
order_by	An optional vector used to determine the order
default	A default value to use if the position does not exist in the input. This is guessed by default for atomic vectors, where a missing value of the appropriate type is return, and for lists, where a <code>NULL</code> is return. For more complicated objects, you'll need to supply this value.

Value

A single value. `[[` is used to do the subsetting.

Examples

```
x <- 1:10
y <- 10:1

last(x)
last(x, y)
```

n_distinct

Efficiently count the number of unique values in a vector.

Description

This is a faster and more concise equivalent of `length(unique(x))`

Usage

```
n_distinct(x)
```

Arguments

x a vector of values

Examples

```
x <- sample(1:10, 1e5, rep = TRUE)
length(unique(x))
n_distinct(x)
```

order_by

A helper function for ordering window function output.

Description

This is a useful function to control the order of window functions in R that don't have a specific ordering parameter. When translated to SQL it will modify the order clause of the OVER function.

Usage

```
order_by(order_by, call)
```

Arguments

order_by a vector to order_by
call a function call to a window function, where the first argument is the vector being operated on

Details

This function works by changing the call to instead call `with_order` with the appropriate arguments.

Examples

```
order_by(10:1, cumsum(1:10))
x <- 10:1
y <- 1:10
order_by(x, cumsum(y))

df <- data.frame(year = 2000:2005, value = (0:5) ^ 2)
scrambled <- df[sample(nrow(df)), ]

wrong <- mutate(scrambled, running = cumsum(value))
arrange(wrong, year)

right <- mutate(scrambled, running = order_by(year, cumsum(value)))
arrange(right, year)
```

ranking

Windowed rank functions.

Description

Six variations on ranking functions, mimicing the ranking functions described in SQL2003. They are currently implemented using the built in rank function, and are provided mainly as a convenience when converting between R and SQL. All ranking functions map smallest inputs to smallest outputs. Use `desc` to reverse the direction..

Usage

```
row_number(x)

ntile(x, n)

min_rank(x)

dense_rank(x)

percent_rank(x)

cume_dist(x)
```

Arguments

x	a vector of values to rank
n	number of groups to split up into.

Details

- row_number: equivalent to `rank(ties.method = "first")`
- min_rank: equivalent to `rank(ties.method = "min")`
- dense_rank: like min_rank, but with no gaps between ranks
- percent_rank: a number between 0 and 1 computed by rescaling min_rank to [0, 1]
- cume_dist: a cumulative distribution function. Proportion of all values less than or equal to the current rank.
- ntile: a rough rank, which breaks the input vector into n buckets.

Examples

```
x <- c(5, 1, 3, 2, 2)
row_number(x)
min_rank(x)
dense_rank(x)
percent_rank(x)
cume_dist(x)

ntile(x, 2)
ntile(runif(100), 10)
```

rbind_list

Efficiently rbind multiple data frames.

Description

This is an efficient version of the common pattern of `do.call(rbind, dfs)` for row-binding many data frames together. It works in the same way as `rbind.fill` but is implemented in C++ so avoids many copies and is much much faster.

Usage

```
rbind_list(...)

rbind_all(dots)
```

Arguments

`dots, ...` list of data frames to combine. With `rbind_all`, they should already be in a list, with `rbind_list` you supply them individually.

Examples

```
one <- mtcars[1:10, ]
two <- mtcars[11:32, ]

rbind_list(one, two)
rbind_all(list(one, two))
```

rowwise

*Group input by rows***Description**

rowwise is used for the results of `do` when you create list-variables. It is also useful to support arbitrary complex operations that need to be applied to each row.

Usage

```
rowwise(data)
```

Arguments

data Input data frame.

Details

Currently rowwise grouping only works with data frames. It's main impact is to allow you to work with list-variables in `summarise` and `mutate` without having to use `[[1]]`. This makes `summarise()` on a rowwise tbl effectively equivalent to `plyr`'s `ldply`.

Examples

```
df <- expand.grid(x = 1:3, y = 3:1)
df %>% rowwise() %>% do(i = seq(.$x, .$y))
.Last.value %>% summarise(n = length(i))
```

sample

*Sample n rows from a table.***Description**

This is a wrapper around `sample.int` to make it easy to select random rows from a table. It currently only works for local tbls.

Usage

```
sample_n(tbl, size, replace = FALSE, weight = NULL, .env = parent.frame())
```

```
sample_frac(tbl, size = 1, replace = FALSE, weight = NULL,
  .env = parent.frame())
```

Arguments

<code>tbl</code>	tbl of data.
<code>size</code>	For <code>sample_n</code> , the number of rows to select. For <code>sample_frac</code> , the fraction of rows to select. If <code>tbl</code> is grouped, <code>size</code> applies to each group.
<code>replace</code>	Sample with or without replacement?
<code>weight</code>	Sampling weights. This expression is evaluated in the context of the data frame. It must return a vector of non-negative numbers the same length as the input. Weights are automatically standardised to sum to 1.
<code>.env</code>	Environment in which to look for non-data names used in <code>weight</code> . Non-default settings for experts only.

Examples

```
by_cyl <- mtcars %>% group_by(cyl)

# Sample fixed number per group
sample_n(mtcars, 10)
sample_n(mtcars, 50, replace = TRUE)
sample_n(mtcars, 10, weight = mpg)

sample_n(by_cyl, 3)
sample_n(by_cyl, 10, replace = TRUE)
sample_n(by_cyl, 3, weight = mpg / mean(mpg))

# Sample fixed fraction per group
# Default is to sample all data = randomly resample rows
sample_frac(mtcars)

sample_frac(mtcars, 0.1)
sample_frac(mtcars, 1.5, replace = TRUE)
sample_frac(mtcars, 0.1, weight = 1 / mpg)

sample_frac(by_cyl, 0.2)
sample_frac(by_cyl, 1, replace = TRUE)
```

select	<i>Select/rename variables by name.</i>
--------	---

Description

`select()` keeps only the variables you mention; `rename()` keeps all variables.

Usage

```
select(.data, ...)
```

```
select_(.data, ..., .dots)
```



```
rename(.data, ...)
```

```
rename_(.data, ...)
```

Arguments

<code>.data</code>	A tbl. All main verbs are S3 generics and provide methods for <code>tbl_df</code> , <code>tbl_dt</code> and <code>tbl_sql</code> .
<code>...</code>	Comma separated list of unquoted expressions. You can treat variable names like they are positions. Use positive values to select variables; use negative values to drop variables.
<code>.dots</code>	Use <code>select_()</code> to do standard evaluation. See <code>vignette("nse")</code> for details

Value

An object of the same class as `.data`.

Data frame row names are silently dropped. To preserve, convert to an explicit variable.

Special functions

As well as using existing functions like `:` and `c`, there are a number of special functions that only work inside `select`

- `starts_with(x, ignore.case = TRUE)`: names starts with `x`
- `ends_with(x, ignore.case = TRUE)`: names ends in `x`
- `contains(x, ignore.case = TRUE)`: selects all variables whose name contains `x`
- `matches(x, ignore.case = TRUE)`: selects all variables whose name matches the regular expression `x`
- `num_range("x", 1:5, width = 2)`: selects all variables (numerically) from `x01` to `x05`.
- `one_of("x", "y", "z")`: selects variables provided in a character vector.
- `everything()`: selects all variables.

To drop variables, use `-`. You can rename variables with named arguments.

See Also

Other `single.table.verbs`: `arrange`, `arrange_`; `filter`, `filter_`; `mutate`, `mutate_`, `transmute`, `transmute_`; `slice`, `slice_`; `summarise`, `summarise_`, `summarize`, `summarize_`

Examples

```
iris <- tbl_df(iris) # so it prints a little nicer
select(iris, starts_with("Petal"))
select(iris, ends_with("Width"))
select(iris, contains("etal"))
select(iris, matches(".t."))
select(iris, Petal.Length, Petal.Width)
```

```

vars <- c("Petal.Length", "Petal.Width")
select(iris, one_of(vars))

df <- as.data.frame(matrix(runif(100), nrow = 10))
df <- tbl_df(df[c(3, 4, 7, 1, 9, 8, 5, 2, 6, 10)])
select(df, V4:V6)
select(df, num_range("V", 4:6))

# Drop variables
select(iris, -starts_with("Petal"))
select(iris, -ends_with("Width"))
select(iris, -contains("etal"))
select(iris, -matches(".t."))
select(iris, -Petal.Length, -Petal.Width)

# Rename variables:
# * select() keeps only the variables you specify
select(iris, petal_length = Petal.Length)
# * rename() keeps all variables
rename(iris, petal_length = Petal.Length)

# Programming with select -----
select_(iris, ~Petal.Length)
select_(iris, "Petal.Length")
select_(iris, lazyeval::interp(~matches(x), x = ".t."))
select_(iris, quote(~Petal.Length), quote(~Petal.Width))
select_(iris, .dots = list(quote(~Petal.Length), quote(~Petal.Width)))

```

setops

Set operations.

Description

These functions override the set functions provided in base to make them generic so that efficient versions for data frames and other tables can be provided. The default methods call the base versions.

Usage

```
intersect(x, y, ...)
```

```
union(x, y, ...)
```

```
setdiff(x, y, ...)
```

```
setequal(x, y, ...)
```

Arguments

`x, y` objects to compare (ignoring order)
`...` other arguments passed on to methods

Examples

```
mtcars$model <- rownames(mtcars)
first <- mtcars[1:20, ]
second <- mtcars[10:32, ]

intersect(first, second)
union(first, second)
setdiff(first, second)
setdiff(second, first)

setequal(mtcars, mtcars[32:1, ])
```

slice	<i>Select rows by position.</i>
-------	---------------------------------

Description

Select rows by position.

Usage

```
slice(.data, ...)

slice_(.data, ..., .dots)
```

Arguments

`.data` A tbl. All main verbs are S3 generics and provide methods for [tbl_df](#), [tbl_dt](#) and [tbl_sql](#).
`...` Integer row values
`.dots` Used to work around non-standard evaluation. See [vignette\("nse"\)](#) for details.

See Also

Other `single.table.verbs`: [arrange](#), [arrange_](#); [filter](#), [filter_](#); [mutate](#), [mutate_](#), [transmute](#), [transmute_](#); [rename](#), [rename_](#), [select](#), [select_](#); [summarise](#), [summarise_](#), [summarize](#), [summarize_](#)

Examples

```

slice(mtcars, 1L)
slice(mtcars, n())
slice(mtcars, 5:n())

by_cyl <- group_by(mtcars, cyl)
slice(by_cyl, 1:2)

```

src_mysql	<i>Connect to mysql/mariadb.</i>
-----------	----------------------------------

Description

Use `src_mysql` to connect to an existing mysql or mariadb database, and `tbl` to connect to tables within that database. If you are running a local mysql database, leave all parameters set as their defaults to connect. If you're connecting to a remote database, ask your database administrator for the values of these variables.

Usage

```

src_mysql(dbname, host = NULL, port = 0L, user = "root", password = "",
  ...)

## S3 method for class 'src_mysql'
tbl(src, from, ...)

```

Arguments

<code>dbname</code>	Database name
<code>host,port</code>	Host name and port number of database
<code>user,password</code>	User name and password. Rather than supplying a username and password here, it's better to save them in <code>my.cnf</code> , as described in MySQL . In that case, supply <code>NULL</code> to both user and password.
<code>...</code>	for the <code>src</code> , other arguments passed on to the underlying database connector, <code>dbConnect</code> . For the <code>tbl</code> , included for compatibility with the generic, but otherwise ignored.
<code>src</code>	a mysql <code>src</code> created with <code>src_mysql</code> .
<code>from</code>	Either a string giving the name of table in database, or sql described a derived table or compound join.

Debugging

To see exactly what SQL is being sent to the database, you see [show_query](#) and [explain](#).

Grouping

Typically you will create a grouped data table is to call the `group_by` method on a mysql tbl: this will take care of capturing the unevaluated expressions for you.

For best performance, the database should have an index on the variables that you are grouping by. Use `explain` to check that the database is using the indexes that you expect.

Output

All data manipulation on SQL tbls are lazy: they will not actually run the query or retrieve the data unless you ask for it: they all return a new `tbl_sql` object. Use `compute` to run the query and save the results in a temporary in the database, or use `collect` to retrieve the results to R.

Note that `do` is not lazy since it must pull the data into R. It returns a `tbl_df` or `grouped_df`, with one column for each grouping variable, and one list column that contains the results of the operation. `do` never simplifies its output.

Query principles

This section attempts to lay out the principles governing the generation of SQL queries from the manipulation verbs. The basic principle is that a sequence of operations should return the same value (modulo class) regardless of where the data is stored.

- `arrange(arrange(df, x), y)` should be equivalent to `arrange(df, y, x)`
- `select(select(df, a:x), n:o)` should be equivalent to `select(df, n:o)`
- `mutate(mutate(df, x2 = x * 2), y2 = y * 2)` should be equivalent to `mutate(df, x2 = x * 2, y2 = y * 2)`
- `filter(filter(df, x == 1), y == 2)` should be equivalent to `filter(df, x == 1, y == 2)`
- `summarise` should return the summarised output with one level of grouping peeled off.

Examples

```
## Not run:
# Connection basics -----
# To connect to a database first create a src:
my_db <- src_mysql(host = "blah.com", user = "hadley",
  password = "pass")
# Then reference a tbl within that src
my_tbl <- tbl(my_db, "my_table")

## End(Not run)

# Here we'll use the Lahman database: to create your own local copy,
# create a local database called "lahman", or tell lahman_mysql() how to
# a database that you can write to

if (!has_lahman("postgres") && has_lahman("mysql")) {
# Methods -----
batting <- tbl(lahman_mysql(), "Batting")
dim(batting)
colnames(batting)
head(batting)
```

```

# Data manipulation verbs -----
filter(batting, yearID > 2005, G > 130)
select(batting, playerID:lgID)
arrange(batting, playerID, desc(yearID))
summarise(batting, G = mean(G), n = n())
mutate(batting, rbi2 = 1.0 * R / AB)

# note that all operations are lazy: they don't do anything until you
# request the data, either by `print()`ing it (which shows the first ten
# rows), by looking at the `head()`, or `collect()` the results locally.

system.time(recent <- filter(batting, yearID > 2010))
system.time(collect(recent))

# Group by operations -----
# To perform operations by group, create a grouped object with group_by
players <- group_by(batting, playerID)
group_size(players)

# MySQL doesn't support windowed functions, which means that only
# grouped summaries are really useful:
summarise(players, mean_g = mean(G), best_ab = max(AB))

# When you group by multiple level, each summarise peels off one level
per_year <- group_by(batting, playerID, yearID)
stints <- summarise(per_year, stints = max(stint))
filter(ungroup(stints), stints > 3)
summarise(stints, max(stints))

# Joins -----
player_info <- select(tbl(lahman_mysql(), "Master"), playerID,
  birthYear)
hof <- select(filter(tbl(lahman_mysql(), "HallOfFame"), inducted == "Y"),
  playerID, votedBy, category)

# Match players and their hall of fame data
inner_join(player_info, hof)
# Keep all players, match hof data where available
left_join(player_info, hof)
# Find only players in hof
semi_join(player_info, hof)
# Find players not in hof
anti_join(player_info, hof)

# Arbitrary SQL -----
# You can also provide sql as is, using the sql function:
batting2008 <- tbl(lahman_mysql(),
  sql("SELECT * FROM Batting WHERE YearID = 2008"))
batting2008
}

```

src_postgres	<i>Connect to postgresql.</i>
--------------	-------------------------------

Description

Use `src_postgres` to connect to an existing postgresql database, and `tbl` to connect to tables within that database. If you are running a local postgresql database, leave all parameters set as their defaults to connect. If you're connecting to a remote database, ask your database administrator for the values of these variables.

Usage

```
src_postgres(dbname = NULL, host = NULL, port = NULL, user = NULL,
             password = NULL, ...)
```

```
## S3 method for class 'src_postgres'
tbl(src, from, ...)
```

Arguments

<code>dbname</code>	Database name
<code>host,port</code>	Host name and port number of database
<code>user,password</code>	User name and password (if needed)
<code>...</code>	for the <code>src</code> , other arguments passed on to the underlying database connector, <code>dbConnect</code> . For the <code>tbl</code> , included for compatibility with the generic, but otherwise ignored.
<code>src</code>	a postgres <code>src</code> created with <code>src_postgres</code> .
<code>from</code>	Either a string giving the name of table in database, or sql described a derived table or compound join.

Debugging

To see exactly what SQL is being sent to the database, you see [show_query](#) and [explain](#).

Grouping

Typically you will create a grouped data table is to call the `group_by` method on a `mysql tbl`: this will take care of capturing the unevaluated expressions for you.

For best performance, the database should have an index on the variables that you are grouping by. Use [explain](#) to check that the database is using the indexes that you expect.

Output

All data manipulation on SQL tbls are lazy: they will not actually run the query or retrieve the data unless you ask for it: they all return a new `tbl_sql` object. Use `compute` to run the query and save the results in a temporary in the database, or use `collect` to retrieve the results to R.

Note that `do` is not lazy since it must pull the data into R. It returns a `tbl_df` or `grouped_df`, with one column for each grouping variable, and one list column that contains the results of the operation. `do` never simplifies its output.

Query principles

This section attempts to lay out the principles governing the generation of SQL queries from the manipulation verbs. The basic principle is that a sequence of operations should return the same value (modulo class) regardless of where the data is stored.

- `arrange(arrange(df, x), y)` should be equivalent to `arrange(df, y, x)`
- `select(select(df, a:x), n:o)` should be equivalent to `select(df, n:o)`
- `mutate(mutate(df, x2 = x * 2), y2 = y * 2)` should be equivalent to `mutate(df, x2 = x * 2, y2 = y * 2)`
- `filter(filter(df, x == 1), y == 2)` should be equivalent to `filter(df, x == 1, y == 2)`
- `summarise` should return the summarised output with one level of grouping peeled off.

Examples

```
## Not run:
# Connection basics -----
# To connect to a database first create a src:
my_db <- src_postgres(host = "blah.com", user = "hadley",
  password = "pass")
# Then reference a tbl within that src
my_tbl <- tbl(my_db, "my_table")

## End(Not run)

# Here we'll use the Lahman database: to create your own local copy,
# create a local database called "lahman", or tell lahman_postgres() how to
# a database that you can write to

if (has_lahman("postgres")) {
# Methods -----
batting <- tbl(lahman_postgres(), "Batting")
dim(batting)
colnames(batting)
head(batting)

# Data manipulation verbs -----
filter(batting, yearID > 2005, G > 130)
select(batting, playerID:lgID)
arrange(batting, playerID, desc(yearID))
summarise(batting, G = mean(G), n = n())
mutate(batting, rbi2 = if(is.null(AB)) 1.0 * R / AB else 0)
```



```

# note that all operations are lazy: they don't do anything until you
# request the data, either by `print()`ing it (which shows the first ten
# rows), by looking at the `head()`, or `collect()` the results locally.

system.time(recent <- filter(batting, yearID > 2010))
system.time(collect(recent))

# Group by operations -----
# To perform operations by group, create a grouped object with group_by
players <- group_by(batting, playerID)
group_size(players)

summarise(players, mean_g = mean(G), best_ab = max(AB))
best_year <- filter(players, AB == max(AB) | G == max(G))
progress <- mutate(players,
  cyear = yearID - min(yearID) + 1,
  ab_rank = rank(desc(AB)),
  cumulative_ab = order_by(yearID, cumsum(AB)))

# When you group by multiple level, each summarise peels off one level
per_year <- group_by(batting, playerID, yearID)
stints <- summarise(per_year, stints = max(stint))
filter(stints, stints > 3)
summarise(stints, max(stints))
mutate(stints, order_by(yearID, cumsum(stints)))

# Joins -----
player_info <- select(tbl(lahman_postgres(), "Master"), playerID, birthYear)
hof <- select(filter(tbl(lahman_postgres(), "HallOfFame"), inducted == "Y"),
  playerID, votedBy, category)

# Match players and their hall of fame data
inner_join(player_info, hof)
# Keep all players, match hof data where available
left_join(player_info, hof)
# Find only players in hof
semi_join(player_info, hof)
# Find players not in hof
anti_join(player_info, hof)

# Arbitrary SQL -----
# You can also provide sql as is, using the sql function:
batting2008 <- tbl(lahman_postgres(),
  sql('SELECT * FROM "Batting" WHERE "yearID" = 2008'))
batting2008
}

```

Description

Use `src_sqlite` to connect to an existing sqlite database, and `tbl` to connect to tables within that database. If you are running a local sqliteql database, leave all parameters set as their defaults to connect. If you're connecting to a remote database, ask your database administrator for the values of these variables.

Usage

```
src_sqlite(path, create = FALSE)

## S3 method for class 'src_sqlite'
tbl(src, from, ...)
```

Arguments

<code>path</code>	Path to SQLite database
<code>create</code>	if FALSE, path must already exist. If TRUE, will create a new SQLite3 database at path.
<code>src</code>	a sqlite src created with <code>src_sqlite</code> .
<code>from</code>	Either a string giving the name of table in database, or sql described a derived table or compound join.
<code>...</code>	Included for compatibility with the generic, but otherwise ignored.

Debugging

To see exactly what SQL is being sent to the database, you see [show_query](#) and [explain](#).

Grouping

Typically you will create a grouped data table is to call the `group_by` method on a `mysql tbl`: this will take care of capturing the unevaluated expressions for you.

For best performance, the database should have an index on the variables that you are grouping by. Use [explain](#) to check that the database is using the indexes that you expect.

Output

All data manipulation on SQL `tbls` are lazy: they will not actually run the query or retrieve the data unless you ask for it: they all return a new `tbl_sql` object. Use [compute](#) to run the query and save the results in a temporary in the database, or use [collect](#) to retrieve the results to R.

Note that `do` is not lazy since it must pull the data into R. It returns a `tbl_df` or `grouped_df`, with one column for each grouping variable, and one list column that contains the results of the operation. `do` never simplifies its output.

Query principles

This section attempts to lay out the principles governing the generation of SQL queries from the manipulation verbs. The basic principle is that a sequence of operations should return the same value (modulo class) regardless of where the data is stored.

- `arrange(arrange(df, x), y)` should be equivalent to `arrange(df, y, x)`
- `select(select(df, a:x), n:o)` should be equivalent to `select(df, n:o)`
- `mutate(mutate(df, x2 = x * 2), y2 = y * 2)` should be equivalent to `mutate(df, x2 = x * 2, y2 = y * 2)`
- `filter(filter(df, x == 1), y == 2)` should be equivalent to `filter(df, x == 1, y == 2)`
- `summarise` should return the summarised output with one level of grouping peeled off.

Examples

```
## Not run:
# Connection basics -----
# To connect to a database first create a src:
my_db <- src_sqlite(path = tempfile(), create = TRUE)
# Then reference a tbl within that src
my_tbl <- tbl(my_db, "my_table")

## End(Not run)

# Here we'll use the Lahman database: to create your own local copy,
# run lahman_sqlite()

if (require("RSQLite") && has_lahman("sqlite")) {
# Methods -----
batting <- tbl(lahman_sqlite(), "Batting")
dim(batting)
colnames(batting)
head(batting)

# Data manipulation verbs -----
filter(batting, yearID > 2005, G > 130)
select(batting, playerID:lgID)
arrange(batting, playerID, desc(yearID))
summarise(batting, G = mean(G), n = n())
mutate(batting, rbi2 = 1.0 * R / AB)

# note that all operations are lazy: they don't do anything until you
# request the data, either by `print()`ing it (which shows the first ten
# rows), by looking at the `head()`, or `collect()` the results locally.

system.time(recent <- filter(batting, yearID > 2010))
system.time(collect(recent))

# Group by operations -----
# To perform operations by group, create a grouped object with group_by
players <- group_by(batting, playerID)
```

```

group_size(players)

# sqlite doesn't support windowed functions, which means that only
# grouped summaries are really useful:
summarise(players, mean_g = mean(G), best_ab = max(AB))

# When you group by multiple level, each summarise peels off one level
per_year <- group_by(batting, playerID, yearID)
stints <- summarise(per_year, stints = max(stint))
filter(ungroup(stints), stints > 3)
summarise(stints, max(stints))

# Joins -----
player_info <- select(tbl(lahman_sqlite(), "Master"), playerID, birthYear)
hof <- select(filter(tbl(lahman_sqlite(), "HallOfFame"), inducted == "Y"),
  playerID, votedBy, category)

# Match players and their hall of fame data
inner_join(player_info, hof)
# Keep all players, match hof data where available
left_join(player_info, hof)
# Find only players in hof
semi_join(player_info, hof)
# Find players not in hof
anti_join(player_info, hof)

# Arbitrary SQL -----
# You can also provide sql as is, using the sql function:
batting2008 <- tbl(lahman_sqlite(),
  sql("SELECT * FROM Batting WHERE YearID = 2008"))
batting2008
}

```

src_tbls

List all tbls provided by a source.

Description

This is a generic method which individual src's will provide methods for. Most methods will not be documented because it's usually pretty obvious what possible results will be.

Usage

```
src_tbls(x)
```

Arguments

x a data src.

summarise	<i>Summarise multiple values to a single value.</i>
-----------	---

Description

Summarise multiple values to a single value.

Usage

```
summarise(.data, ...)  
  
summarise_(.data, ..., .dots)  
  
summarize(.data, ...)  
  
summarize_(.data, ..., .dots)
```

Arguments

.data	A tbl. All main verbs are S3 generics and provide methods for tbl_df , tbl_dt and tbl_sql .
...	Name-value pairs of summary functions like min() , mean() , max() etc.
.dots	Used to work around non-standard evaluation. See vignette("nse") for details.

Value

An object of the same class as .data. One grouping level will be dropped.

Data frame row names are silently dropped. To preserve, convert to an explicit variable.

See Also

Other `single.table.verbs`: [arrange](#), [arrange_](#); [filter](#), [filter_](#); [mutate](#), [mutate_](#), [transmute](#), [transmute_](#); [rename](#), [rename_](#), [select](#), [select_](#); [slice](#), [slice_](#)

Examples

```
summarise(mtcars, mean(displacement))  
summarise(group_by(mtcars, cyl), mean(displacement))  
  
summarise(group_by(mtcars, cyl), m = mean(displacement), sd = sd(displacement))
```

summarise_each	<i>Summarise and mutate multiple columns.</i>
----------------	---

Description

Apply one or more functions to one or more columns. Grouping variables are always excluded from modification.

Usage

```
summarise_each(tbl, funs, ...)

summarise_each_(tbl, funs, vars)

mutate_each(tbl, funs, ...)

mutate_each_(tbl, funs, vars)
```

Arguments

tbl	a tbl
funs	List of function calls, generated by funs , or a character vector of function names.
vars, ...	Variables to include/exclude in mutate/summarise. You can use same specifications as in select . If missing, defaults to all non-grouping variables. For standard evaluation versions (ending in <code>_q</code>) these can be either a list of expressions or a character vector.

Examples

```
# One function
by_species <- iris %>% group_by(Species)
by_species %>% summarise_each(funs(length))
by_species %>% summarise_each(funs(mean))
by_species %>% summarise_each(funs(mean), Petal.Width)
by_species %>% summarise_each(funs(mean), matches("Width"))

by_species %>% mutate_each(funs(half = . / 2))
by_species %>% mutate_each(funs(min_rank))

# Two functions
by_species %>% summarise_each(funs(min, max))
by_species %>% summarise_each(funs(min, max), Petal.Width, Sepal.Width)
by_species %>% summarise_each(funs(min, max), matches("Width"))

# Alternative function specification
iris %>% summarise_each(funs(ul = length(unique(.))))
by_species %>% summarise_each(funs(ul = length(unique(.))))
```

```
by_species %>% summarise_each(c("min", "max"))

# Alternative variable specification
summarise_each_(iris, funs(max), names(iris)[-5])
summarise_each_(iris, funs(max), list(quote(-Species)))
```

tally

Counts/tally observations by group.

Description

tally is a convenient wrapper for summarise that will either call `n` or `sum(n)` depending on whether you're tallying for the first time, or re-tallying. `tally()` is similar, but also does the `group_by` for you.

Usage

```
tally(x, wt, sort = FALSE)

count(x, ..., wt = NULL, sort = FALSE)
```

Arguments

x	a tbl to tally/count.
wt	(Optional) If not specified, will tally the number of rows. If specified, will perform a "weighted" tally but summing over the specified variable.
sort	if TRUE will sort output in descending order of n
...	Variables to group by.

Examples

```
if (require("Lahman")) {
  batting_tbl <- tbl_df(Batting)
  tally(group_by(batting_tbl, yearID))
  tally(group_by(batting_tbl, yearID), sort = TRUE)

  # Multiple tallys progressively roll up the groups
  plays_by_year <- tally(group_by(batting_tbl, playerID, stint), sort = TRUE)
  tally(plays_by_year, sort = TRUE)
  tally(tally(plays_by_year))

  # This looks a little nicer if you use the infix %>% operator
  batting_tbl %>% group_by(playerID) %>% tally(sort = TRUE)

  # count is even more succinct - it also does the grouping for you
  batting_tbl %>% count(playerID)
  batting_tbl %>% count(playerID, wt = G)
  batting_tbl %>% count(playerID, wt = G, sort = TRUE)
}
```

tbl	<i>Create a table from a data source</i>
-----	--

Description

This is a generic method that dispatches based on the first argument.

Usage

```
tbl(src, ...)
```

```
is.tbl(x)
```

```
as.tbl(x, ...)
```

Arguments

src	A data source
...	Other arguments passed on to the individual methods
x	an object to coerce to a tbl

tbl_cube	<i>A data cube tbl.</i>
----------	-------------------------

Description

An cube tbl stores data in a compact array format where dimension names are not needlessly repeated. They are particularly appropriate for experimental data where all combinations of factors are tried (e.g. complete factorial designs), or for storing the result of aggregations. Compared to data frames, they will occupy much less memory when variables are crossed, not nested.

Usage

```
tbl_cube(dimensions, measures)
```

Arguments

dimensions	A named list of vectors. A dimension is a variable whose values are known before the experiment is conducted; they are fixed by design (in reshape2 they are known as id variables). tbl_cubes are dense which means that almost every combination of the dimensions should have associated measurements: missing values require an explicit NA, so if the variables are nested, not crossed, the majority of the data structure will be empty. Dimensions are typically, but not always, categorical variables.
------------	---

measures A named list of arrays. A measure is something that is actually measured, and is not known in advance. The dimension of each array should be the same as the length of the dimensions. Measures are typically, but not always, continuous values.

Details

tbl_cube support is currently experimental and little performance optimisation has been done, but you may find them useful if your data already comes in this form, or you struggle with the memory overhead of the sparse/crossed of data frames. There is no supported for hierarchical indices (although I think that would be a relatively straightforward extension to storing data frames for indices rather than vectors).

Implementation

Manipulation functions:

- `select (M)`
- `summarise (M)`, corresponds to roll-up, but rather more limited since there are no hierarchies.
- `filter (D)`, corresponds to slice/dice.
- `mutate (M)` is not implemented, but should be relatively straightforward given the implementation of `summarise`.
- `arrange (D?)` Not implemented: not obvious how much sense it would make

Joins: not implemented. See `vignettes/joins.graffle` for ideas. Probably straightforward if you get the indexes right, and that's probably some straightforward array/tensor operation.

See Also

[as.tbl_cube](#) for ways of coercing existing data structures into a `tbl_cube`.

Examples

```
# The built in nasa dataset records meterological data (temperature,
# cloud cover, ozone etc) for a 4d spatio-temporal dataset (lat, long,
# month and year)
nasa
head(as.data.frame(nasa))

titanic <- as.tbl_cube(Titanic)
head(as.data.frame(titanic))

admit <- as.tbl_cube(UCBAdmissions)
head(as.data.frame(admit))

as.tbl_cube(esoph, dim_names = 1:3)

# Some manipulation examples with the NASA dataset -----

# select() operates only on measures: it doesn't affect dimensions in any way
```

```

select(nasa, cloudhigh:cloudmid)
select(nasa, matches("temp"))

# filter() operates only on dimensions
filter(nasa, lat > 0, year == 2000)
# Each component can only refer to one dimensions, ensuring that you always
# create a rectangular subset
## Not run: filter(nasa, lat > long)

# Arrange is meaningless for tbl_cubes

by_loc <- group_by(nasa, lat, long)
summarise(by_loc, pressure = max(pressure), temp = mean(temperature))

```

tbl_df	<i>Create a data frame tbl.</i>
--------	---------------------------------

Description

A data frame `tbl` wraps a local data frame. The main advantage to using a `tbl_df` over a regular data frame is the printing: `tbl` objects only print a few rows and all the columns that fit on one screen, describing the rest of it as text.

Usage

```
tbl_df(data)
```

Arguments

`data` a data frame

Methods

`tbl_df` implements two important base methods:

print Only prints the first 10 rows, and the columns that fit on screen

[Never simplifies (drops), so always returns `data.frame`

Examples

```

ds <- tbl_df(mtcars)
ds
as.data.frame(ds)

if (require("Lahman") && packageVersion("Lahman") >= "3.0.1") {
  batting <- tbl_df(Batting)
  dim(batting)
  colnames(batting)
  head(batting)
}

```

```

# Data manipulation verbs -----
filter(batting, yearID > 2005, G > 130)
select(batting, playerID:lgID)
arrange(batting, playerID, desc(yearID))
summarise(batting, G = mean(G), n = n())
mutate(batting, rbi2 = if(is.null(AB)) 1.0 * R / AB else 0)

# Group by operations -----
# To perform operations by group, create a grouped object with group_by
players <- group_by(batting, playerID)
head(group_size(players), 100)

summarise(players, mean_g = mean(G), best_ab = max(AB))
best_year <- filter(players, AB == max(AB) | G == max(G))
progress <- mutate(players, cyear = yearID - min(yearID) + 1,
  rank(desc(AB)), cumsum(AB))

# When you group by multiple level, each summarise peels off one level
per_year <- group_by(batting, playerID, yearID)
stints <- summarise(per_year, stints = max(stint))
# filter(stints, stints > 3)
# summarise(stints, max(stints))
# mutate(stints, cumsum(stints))

# Joins -----
player_info <- select(tbl_df(Master), playerID, birthYear)
hof <- select(filter(tbl_df(HallOfFame), inducted == "Y"),
  playerID, votedBy, category)

# Match players and their hall of fame data
inner_join(player_info, hof)
# Keep all players, match hof data where available
left_join(player_info, hof)
# Find only players in hof
semi_join(player_info, hof)
# Find players not in hof
anti_join(player_info, hof)
}

```

tbl_dt

*Create a data table tbl.***Description**

A data table tbl wraps a local data table.

Usage

```
tbl_dt(data, copy = TRUE)
```

Arguments

data	a data table
copy	If the input is a data.table, copy it?

Examples

```

if (require("data.table")) {
  ds <- tbl_dt(mtcars)
  ds
  as.data.table(ds)
  as.tbl(mtcars)
}

if (require("data.table") && require("nycflights13")) {
  flights2 <- tbl_dt(flights)
  flights2 %>% filter(month == 1, day == 1, dest == "DFW")
  flights2 %>% select(year:day)
  flights2 %>% rename(Year = year)
  flights2 %>%
    summarise(
      delay = mean(arr_delay, na.rm = TRUE),
      n = length(arr_delay)
    )
  flights2 %>%
    mutate(gained = arr_delay - dep_delay) %>%
    select(ends_with("delay"), gained)
  flights2 %>%
    arrange(dest, desc(arr_delay))

  by_dest <- group_by(flights2, dest)

  filter(by_dest, arr_delay == max(arr_delay, na.rm = TRUE))
  summarise(by_dest, arr = mean(arr_delay, na.rm = TRUE))

  # Normalise arrival and departure delays by airport
  by_dest %>%
    mutate(arr_z = scale(arr_delay), dep_z = scale(dep_delay)) %>%
    select(starts_with("arr"), starts_with("dep"))

  arrange(by_dest, desc(arr_delay))
  select(by_dest, -(day:tailnum))
  rename(by_dest, Year = year)

  # All manip functions preserve grouping structure, except for summarise
  # which removes a grouping level
  by_day <- group_by(flights2, year, month, day)
  by_month <- summarise(by_day, delayed = sum(arr_delay > 0, na.rm = TRUE))
  by_month
  summarise(by_month, delayed = sum(delayed))

  # You can also manually ungroup:
  ungroup(by_day)

```

```
}  
  

```

tbl_vars	List variables provided by a tbl.
----------	-----------------------------------

Description

List variables provided by a tbl.

Usage

```
tbl_vars(x)
```

Arguments

x A tbl object

top_n	Select top n rows (by value).
-------	-------------------------------

Description

This is a convenient wrapper that uses [filter](#) and [min_rank](#) to select the top n entries in each group, ordered by wt.

Usage

```
top_n(x, n, wt)
```

Arguments

x a [tbl](#) to filter

n number of rows to return. If x is grouped, this is the number of rows per group. May include more than n if there are ties.

wt (Optional). The variable to use for ordering. If not specified, defaults to the last variable in the tbl.

Examples

```
# Find 10 players with most games
if (require("Lahman")) {
  players <- group_by(tbl_df(Batting), playerID)
  games <- tally(players, G)
  top_n(games, 10, n)

  # A little nicer with %>%
  tbl_df(Batting) %>%
    group_by(playerID) %>%
    tally(G) %>%
    top_n(10)

  # Find year with most games for each player
  tbl_df(Batting) %>% group_by(playerID) %>% top_n(1, G)
}
```

translate_sql	<i>Translate an expression to sql.</i>
---------------	--

Description

Translate an expression to sql.

Usage

```
translate_sql(..., tbl = NULL, env = parent.frame(), variant = NULL,
  window = FALSE)

translate_sql_q(expr, tbl = NULL, env = parent.frame(), variant = NULL,
  window = FALSE)
```

Arguments

...	unevaluated expression to translate
tbl	An optional tbl . If supplied, will be used to automatically figure out the SQL variant to use.
env	environment in which to evaluate expression.
variant	used to override default variant provided by source useful for testing/examples
window	If variant not supplied, used to determine whether the variant is window based or not.
expr	list of quoted objects to translate

Base translation

The base translator, `base_sql`, provides custom mappings for `!` (to `NOT`), `&&` and `&` to `AND`, `||` and `|` to `OR`, `^` to `POWER`, `%>%` to `%`, `ceiling` to `CEIL`, `mean` to `AVG`, `var` to `VARIANCE`, `tolower` to `LOWER`, `toupper` to `UPPER` and `nchar` to `length`.

`c` and `:` keep their usual R behaviour so you can easily create vectors that are passed to sql.

All other functions will be preserved as is. R's infix functions (e.g. `%like%`) will be converted to their sql equivalents (e.g. `LIKE`). You can use this to access SQL string concatenation: `||` is mapped to `OR`, but `%||%` is mapped to `||`. To suppress this behaviour, and force errors immediately when `dplyr` doesn't know how to translate a function it encounters, using `set the dplyr.strict_sql` option to `TRUE`.

You can also use `sql` to insert a raw sql string.

SQLite translation

The SQLite variant currently only adds one additional function: a mapping from `sd` to the SQL aggregation function `stdev`.

Examples

```
# Regular maths is translated in a very straightforward way
translate_sql(x + 1)
translate_sql(sin(x) + tan(y))

# Logical operators are converted to their sql equivalents
translate_sql(x < 5 & !(y >= 5))

# If is translated into select case
translate_sql(if (x > 5) "big" else "small")

# Infix functions are passed onto SQL with % removed
translate_sql(first %like% "Had*")
translate_sql(first %is% NULL)
translate_sql(first %in% c("John", "Roger", "Robert"))

# Note that variable names will be escaped if needed
translate_sql(like == 7)

# And be careful if you really want integers
translate_sql(x == 1)
translate_sql(x == 1L)

# If you have an already quoted object, use translate_sql_q:
x <- quote(y + 1 / sin(t))
translate_sql(x)
translate_sql_q(list(x))

# Translation with data source -----

flights <- tbl(nycflights13_postgres(), "flights")
# Note distinction between integers and reals
```

```

translate_sql(month == 1, tbl = flights)
translate_sql(month == 1L, tbl = flights)

# Know how to translate most simple mathematical expressions
translate_sql(month %in% 1:3, tbl = flights)
translate_sql(month >= 1L & month <= 3L, tbl = flights)
translate_sql((month >= 1L & month <= 3L) | carrier == "AA", tbl = flights)

# Some R functions don't have equivalents in SQL: where possible they
# will be translated to the equivalent
translate_sql(xor(month <= 3L, carrier == "AA"), tbl = flights)

# Local variables will be automatically inserted into the SQL
x <- 5L
translate_sql(month == x, tbl = flights)

# By default all computation will happen in sql
translate_sql(month < 1 + 1, source = flights)
# Use local to force local evaluation
translate_sql(month < local(1 + 1), source = flights)

# This is also needed if you call a local function:
inc <- function(x) x + 1
translate_sql(month == inc(x), source = flights)
translate_sql(month == local(inc(x)), source = flights)

# Windowed translation -----
planes <- arrange(group_by(flights, tailnum), desc(DepTime))

translate_sql(dep_time > mean(dep_time), tbl = planes, window = TRUE)
translate_sql(dep_time == min(dep_time), tbl = planes, window = TRUE)

translate_sql(rank(), tbl = planes, window = TRUE)
translate_sql(rank(dep_time), tbl = planes, window = TRUE)
translate_sql(ntile(dep_time, 2L), tbl = planes, window = TRUE)
translate_sql(lead(dep_time, 2L), tbl = planes, window = TRUE)
translate_sql(cumsum(dep_time), tbl = planes, window = TRUE)
translate_sql(order_by(dep_time, cumsum(dep_time)), tbl = planes, window = TRUE)

```


Index

*Topic **debugging**

- failwith, [19](#)
- .datatable.aware (tbl_dt), [59](#)
- [[, [35](#)
- %.%(chain), [9](#)
- %>%(chain), [9](#)
- all.equal.tbl_df, [3](#)
- all.equal.tbl_dt (all.equal.tbl_df), [3](#)
- anti_join (join), [25](#)
- anti_join.data.table (join.tbl_dt), [28](#)
- anti_join.tbl_df (join.tbl_df), [27](#)
- anti_join.tbl_sql (join.tbl_sql), [29](#)
- arrange, [4](#), [16](#), [20](#), [33](#), [41](#), [43](#), [53](#)
- arrange_, [20](#), [33](#), [41](#), [43](#), [53](#)
- arrange_ (arrange), [4](#)
- as.tbl (tbl), [56](#)
- as.tbl_cube, [5](#), [57](#)
- bench_compare, [6](#)
- bench_tbls (bench_compare), [6](#)
- between, [7](#)
- build_sql, [8](#)
- chain, [9](#)
- chain_q (chain), [9](#)
- changes (location), [32](#)
- collapse (compute), [10](#)
- collect, [45](#), [48](#), [50](#)
- collect (compute), [10](#)
- compare_tbls (bench_compare), [6](#)
- compute, [10](#), [45](#), [48](#), [50](#)
- copy_to, [11](#), [11](#)
- copy_to.src_sql, [12](#)
- count (tally), [55](#)
- cumall, [13](#)
- cumany (cumall), [13](#)
- cume_dist (ranking), [37](#)
- cummean (cumall), [13](#)
- data.frame, [14](#)
- data_frame, [13](#)
- data_frame_ (data_frame), [13](#)
- dense_rank (ranking), [37](#)
- desc, [4](#), [14](#), [37](#)
- dimnames, [6](#)
- distinct, [15](#)
- distinct_ (distinct), [15](#)
- do, [16](#), [39](#)
- dplyr, [17](#)
- dplyr-package (dplyr), [17](#)
- eval_tbls (bench_compare), [6](#)
- explain, [18](#), [44](#), [45](#), [47](#), [50](#)
- failwith, [19](#)
- filter, [5](#), [16](#), [20](#), [33](#), [34](#), [41](#), [43](#), [53](#), [61](#)
- filter_, [5](#), [33](#), [41](#), [43](#), [53](#)
- filter_ (filter), [20](#)
- first (nth), [35](#)
- funs, [21](#), [54](#)
- funs_ (funs), [21](#)
- glimpse, [21](#)
- group_by, [23](#), [55](#)
- group_by_ (group_by), [23](#)
- group_size, [25](#)
- grouped_df, [24](#), [45](#), [48](#), [50](#)
- grouped_dt, [22](#), [24](#)
- groups, [23](#), [24](#)
- ident, [8](#)
- inner_join (join), [25](#)
- inner_join.data.table (join.tbl_dt), [28](#)
- inner_join.tbl_df (join.tbl_df), [27](#)
- inner_join.tbl_sql (join.tbl_sql), [29](#)
- intersect (setops), [42](#)
- is.grouped_dt (grouped_dt), [22](#)
- is.tbl (tbl), [56](#)
- isTRUE, [4](#)
- join, [25](#), [27–29](#)

join.tbl_df, 27
 join.tbl_dt, 28
 join.tbl_sql, 29

 lag.default (lead-lag), 31
 last (nth), 35
 lazy_dots, 14
 lead (lead-lag), 31
 lead-lag, 31
 left_join (join), 25
 left_join.data.table (join.tbl_dt), 28
 left_join.tbl_df (join.tbl_df), 27
 left_join.tbl_sql (join.tbl_sql), 29
 location, 32

 max, 53
 mean, 53
 microbenchmark, 6
 min, 53
 min_rank, 61
 min_rank (ranking), 37
 mutate, 5, 16, 20, 33, 34, 39, 41, 43, 53
 mutate_, 5, 20, 41, 43, 53
 mutate_ (mutate), 33
 mutate_each (summarise_each), 54
 mutate_each_ (summarise_each), 54
 mutate_each_q (summarise_each), 54
 MySQL, 44

 n, 34, 55
 n_distinct, 36
 n_groups (group_size), 25
 nasa, 34
 nth, 35
 ntile (ranking), 37

 order_by, 36

 percent_rank (ranking), 37
 print, 18

 ranking, 37
 rbind.fill, 38
 rbind_all (rbind_list), 38
 rbind_list, 38
 regroup (group_by), 23
 rename, 5, 20, 33, 43, 53
 rename (select), 40
 rename_, 5, 20, 33, 43, 53
 rename_ (select), 40

 row_number (ranking), 37
 rowwise, 16, 39

 sample, 39
 sample.int, 39
 sample_frac (sample), 39
 sample_n (sample), 39
 select, 5, 16, 20, 33, 40, 43, 53, 54
 select_, 5, 20, 33, 43, 53
 select_ (select), 40
 semi_join (join), 25
 semi_join.data.table (join.tbl_dt), 28
 semi_join.tbl_df (join.tbl_df), 27
 semi_join.tbl_sql (join.tbl_sql), 29
 setdiff (setops), 42
 setequal (setops), 42
 setops, 42
 show_query, 44, 47, 50
 show_query (explain), 18
 slice, 5, 20, 33, 41, 43, 53
 slice_, 5, 20, 33, 41, 53
 slice_ (slice), 43
 sql, 8, 44, 47, 50
 src_local, 6
 src_mysql, 24, 44
 src_postgres, 24, 47
 src_sqlite, 24, 49
 src_tbls, 52
 str, 18, 21
 sum, 55
 summarise, 5, 16, 20, 33, 34, 39, 41, 43, 53
 summarise_, 5, 20, 33, 41, 43
 summarise_ (summarise), 53
 summarise_each, 54
 summarise_each_ (summarise_each), 54
 summarise_each_q (summarise_each), 54
 summarize, 5, 20, 33, 41, 43
 summarize (summarise), 53
 summarize_, 5, 20, 33, 41, 43
 summarize_ (summarise), 53

 tally, 55
 tbl, 6, 12, 23, 55, 56, 61, 62
 tbl.src_mysql (src_mysql), 44
 tbl.src_postgres (src_postgres), 47
 tbl.src_sqlite (src_sqlite), 49
 tbl_cube, 34, 56
 tbl_df, 4, 10, 20, 27, 33, 41, 43, 45, 48, 50, 53, 58

tbl_dt, [4](#), [20](#), [33](#), [41](#), [43](#), [53](#), [59](#)
tbl_sql, [4](#), [20](#), [33](#), [41](#), [43](#), [45](#), [48](#), [50](#), [53](#)
tbl_vars, [61](#)
top_n, [61](#)
translate_sql, [62](#)
translate_sql_q(translate_sql), [62](#)
transmute, [5](#), [20](#), [41](#), [43](#), [53](#)
transmute(mutate), [33](#)
transmute_, [5](#), [20](#), [41](#), [43](#), [53](#)
transmute_(mutate), [33](#)
try_default, [19](#)

ungroup, [24](#)
ungroup(groups), [23](#)
union(setops), [42](#)
unique, [15](#)

with_order, [37](#)