

Package ‘Rdatasets’

June 6, 2025

Type Package

Title Access Datasets from the Rdatasets Archive

Version 0.0.1

Date 2025-05-31

Description Download and access datasets from the Rdatasets archive (<https://vincentarelbundock.github.io/Rdatasets/>). The package provides functions to search, download, and view documentation for thousands of datasets from various R packages, available in both CSV and Parquet formats for efficient access.

License GPL (>= 3)

URL <https://vincentarelbundock.github.io/Rdatasetspkg/>,
<https://vincentarelbundock.github.io/Rdatasets/>

BugReports <https://github.com/vincentarelbundock/Rdatasetspkg/issues>

Imports utils

Suggests nanoparquet, tinytable, tinytest, rstudioapi, tibble,
data.table

Encoding UTF-8

RoxygenNote 7.3.2

NeedsCompilation no

Author Vincent Arel-Bundock [aut, cre, cph] (ORCID:
<https://orcid.org/0000-0003-2042-7063>)

Maintainer Vincent Arel-Bundock <vincent.arel-bundock@umontreal.ca>

Repository CRAN

Date/Publication 2025-06-06 12:50:04 UTC

Contents

rddata	2
rddocs	3
rdindex	4
rdsearch	5

rddata *Download and Read Datasets from Rdatasets*

Description

Downloads a dataset from the Rdatasets archive and returns it as a data frame.

<https://vincentarelbundock.github.io/Rdatasets/>

Usage

```
rddata(dataset, package = NULL)
```

Arguments

dataset	String. Name of the dataset to download from the Rdatasets archive. Use <code>rdsearch()</code> to search available datasets.
package	String. Package name that originally published the data. If <code>NULL</code> , the function will attempt to automatically detect the package by searching for an exact match in the Rdatasets index.

Details

If the `nanoparquet` package is installed, `rddata()` will use the Parquet format, which is faster and uses less bandwidth to download. If `nanoparquet` is not available, the function automatically falls back to CSV format using base R functionality.

Value

A data frame containing the dataset. The columns and rows vary based on the dataset.

Global Options

The following global options control package behavior:

- `Rdatasets_cache`: Logical
 - Whether to cache downloaded data and index for faster subsequent access. Default: `TRUE`. Please keep this option `TRUE` as it makes repeated access faster and avoids overloading the Rdatasets server. Only set to `FALSE` if local memory is severely limited.
 - Ex: `options(Rdatasets_cache = TRUE)`
- `Rdatasets_class`: String
 - Output class of the returned data. One of `"data.frame"` (default), `"tibble"`, or `"data.table"`. Default: `"data.frame"`. Requires the respective packages to be installed for `"tibble"` or `"data.table"` formats.
 - Ex: `options(Rdatasets_class = "tibble")`
- `Rdataset_path`: String.

- Base URL for the Rdatasets archive. Default: "https://vincentarelbundock.github.io/Rdatasets/". Advanced users can set this to use a different mirror or local copy.
- Ex: `options(Rdataset_path = "https://vincentarelbundock.github.io/Rdatasets/")`

Examples

```
dat <- rddata("Titanic", "Stat2Data")
head(dat)
```

rddocs

Open Dataset Documentation

Description

Opens the documentation for a dataset from Rdatasets as an HTML page using `getOption("viewer")` or the Rstudio viewer.

Usage

```
rddocs(dataset, package = NULL)
```

Arguments

dataset	String. Name of the dataset to download from the Rdatasets archive. Use <code>rdsearch()</code> to search available datasets.
package	String. Package name that originally published the data. If NULL, the function will attempt to automatically detect the package by searching for an exact match in the Rdatasets index.

Details

The function attempts to open the documentation in the following order:

1. RStudio's built-in viewer (if `rstudioapi` is available)
2. The viewer specified in `getOption("viewer")`
3. The default browser specified in `getOption("browser")`

To control which viewer is used, you can set the following options:

- `options(viewer = function(url) { ... })` - Set a custom viewer function
- `options(browser = "firefox")` - Set the default browser (used as fallback)

If no viewer is available, the function will stop with an error message.

Value

Invisibly returns NULL. The function's primary purpose is to open the dataset documentation in a viewer window.

Global Options

The following global options control package behavior:

- `Rdatasets_cache`: Logical
 - Whether to cache downloaded data and index for faster subsequent access. Default: TRUE. Please keep this option TRUE as it makes repeated access faster and avoids overloading the Rdatasets server. Only set to FALSE if local memory is severely limited.
 - Ex: `options(Rdatasets_cache = TRUE)`
- `Rdatasets_class`: String
 - Output class of the returned data. One of "data.frame" (default), "tibble", or "data.table". Default: "data.frame". Requires the respective packages to be installed for "tibble" or "data.table" formats.
 - Ex: `options(Rdatasets_class = "tibble")`
- `Rdataset_path`: String.
 - Base URL for the Rdatasets archive. Default: "https://vincentarelbundock.github.io/Rdatasets/". Advanced users can set this to use a different mirror or local copy.
 - Ex: `options(Rdataset_path = "https://vincentarelbundock.github.io/Rdatasets/")`

Examples

```
rddocs(dataset = "Titanic", package = "Stat2Data")
rddocs("iris", "datasets")
```

rdindex

Get Rdatasets Index

Description

Downloads and returns the complete Rdatasets index as a data frame.

Usage

```
rdindex()
```

Value

A data frame containing all available datasets from Rdatasets with the following columns:

- `Package`: Character. The name of the R package that contains the dataset
- `Dataset`: Character. The name of the dataset
- `Title`: Character. A descriptive title for the dataset
- `Rows`: Integer. Number of rows in the dataset
- `Cols`: Integer. Number of columns in the dataset

- `n_binary`: Integer. Number of binary variables in the dataset
- `n_character`: Integer. Number of character variables in the dataset
- `n_factor`: Integer. Number of factor variables in the dataset
- `n_logical`: Integer. Number of logical variables in the dataset
- `n_numeric`: Integer. Number of numeric variables in the dataset
- `CSV`: Character. URL to download the dataset in CSV format
- `Doc`: Character. URL to the dataset's documentation

Global Options

The following global options control package behavior:

- `Rdatasets_cache`: Logical
 - Whether to cache downloaded data and index for faster subsequent access. Default: `TRUE`. Please keep this option `TRUE` as it makes repeated access faster and avoids overloading the Rdatasets server. Only set to `FALSE` if local memory is severely limited.
 - Ex: `'options(Rdatasets_cache = TRUE)'`
- `Rdatasets_class`: String
 - Output class of the returned data. One of `"data.frame"` (default), `"tibble"`, or `"data.table"`. Default: `"data.frame"`. Requires the respective packages to be installed for `"tibble"` or `"data.table"` formats.
 - Ex: `options(Rdatasets_class = "tibble")`
- `Rdataset_path`: String.
 - Base URL for the Rdatasets archive. Default: `"https://vincentarelbundock.github.io/Rdatasets/"`. Advanced users can set this to use a different mirror or local copy.
 - Ex: `options(Rdataset_path = "https://vincentarelbundock.github.io/Rdatasets/")`

Examples

```
idx <- rdindex()
head(idx)
```

rdsearch

Search Available Datasets

Description

Search available datasets from the Rdatasets archive by regular expression.

Usage

```
rdsearch(
  pattern,
  field = NULL,
  fixed = FALSE,
  perl = FALSE,
  ignore.case = FALSE
)
```

Arguments

pattern	String. Search pattern. Can be a regular expression or literal string depending on the fixed argument.
field	String. Which field to search in. One of "package", "dataset", "title". If NULL (default), searches in all three fields.
fixed	logical. If TRUE, pattern is a string to be matched as is. Overrides all conflicting arguments.
perl	logical. Should Perl-compatible regexps be used?
ignore.case	logical. if FALSE, the pattern matching is <i>case sensitive</i> and if TRUE, case is ignored during matching.

Value

A data frame containing matching datasets with the following columns:

- Package: Character. The name of the R package that contains the dataset
- Dataset: Character. The name of the dataset
- Title: Character. A descriptive title for the dataset
- Rows: Integer. Number of rows in the dataset
- Cols: Integer. Number of columns in the dataset
- n_binary: Integer. Number of binary variables in the dataset
- n_character: Integer. Number of character variables in the dataset
- n_factor: Integer. Number of factor variables in the dataset
- n_logical: Integer. Number of logical variables in the dataset
- n_numeric: Integer. Number of numeric variables in the dataset
- CSV: Character. URL to download the dataset in CSV format
- Doc: Character. URL to the dataset's documentation

Global Options

The following global options control package behavior:

- Rdatasets_cache: Logical

- Whether to cache downloaded data and index for faster subsequent access. Default: TRUE. Please keep this option TRUE as it makes repeated access faster and avoids overloading the Rdatasets server. Only set to FALSE if local memory is severely limited.
- Ex: `options(Rdatasets_cache = TRUE)`
- `Rdatasets_class`: String
 - Output class of the returned data. One of "data.frame" (default), "tibble", or "data.table". Default: "data.frame". Requires the respective packages to be installed for "tibble" or "data.table" formats.
 - Ex: `options(Rdatasets_class = "tibble")`
- `Rdataset_path`: String.
 - Base URL for the Rdatasets archive. Default: "https://vincentarelbundock.github.io/Rdatasets/". Advanced users can set this to use a different mirror or local copy.
 - Ex: `options(Rdataset_path = "https://vincentarelbundock.github.io/Rdatasets/")`

Examples

```
# Search all fields (default behavior)
rdsearch("iris")

# Case-insensitive search
rdsearch("(?i)titanic")

# Search only in package names
rdsearch("datasets", field = "package")

# Search only in dataset names
rdsearch("iris", field = "dataset")

# Search only in titles
rdsearch("Edgar Anderson", field = "title")
```

Index

rddata, 2
rddocs, 3
rdindex, 4
rdsearch, 5