Package 'cstidy'

May 24, 2023

Title Helpful Functions for Cleaning Surveillance Data

Version 2023.5.24

Description
Helpful functions for the cleaning and manipulation of surveillance data, especially with regards to the creation and validation of panel data from individual level surveillance data.

Depends R (>= 3.5.0)

License MIT + file LICENSE

URL https://www.csids.no/cstidy/, https://github.com/csids/cstidy

BugReports https://github.com/csids/cstidy/issues

Encoding UTF-8

LazyData true

Imports data.table, magrittr, ggplot2, csdata, cstime, crayon, digest, stringr, methods

Suggests testthat, knitr, rmarkdown, rstudioapi, glue, gt, dplyr, purrr

RoxygenNote 7.2.3

VignetteBuilder knitr

NeedsCompilation no

Author Richard Aubrey White [aut, cre] (<https://orcid.org/0000-0002-6747-1726>), CSIDS [cph]

Maintainer Richard Aubrey White <hello@rwhite.no>

Repository CRAN

Date/Publication 2023-05-24 07:40:02 UTC

R topics documented:

`csdb_validator_field_contents_csfmt_rts_data_v1`
Field contents validator (csfmt_rts_data_v1) An example (schema) validator of database data used in csfmt_rts_data_v1

Description

Field contents validator (csfmt_rts_data_v1) An example (schema) validator of database data used in csfmt_rts_data_v1

Usage

```
csdb_validator_field_contents_csfmt_rts_data_v1(data)
```

Arguments

`data`	data passed to schema

Value

Boolean, corresponding to where or not the validator is passed.

`csdb_validator_field_types_csfmt_rts_data_v1`
Field types validator (csfmt_rts_data_v1) An example (schema) validator of field_types used in csfmt_rts_data_v1

Description

Field types validator (csfmt_rts_data_v1) An example (schema) validator of field_types used in csfmt_rts_data_v1

Usage

```
csdb_validator_field_types_csfmt_rts_data_v1(db_field_types)
```

Arguments

db_field_types	db_field_types passed to schema

Value

Boolean, corresponding to where or not the validator is passed.

`expand_time_to` *Expand time to*

Description

Attempts to expand the dataset to include more time

A time series is defined as a unique combination of:

- granularity_time
- granularity_geo
- country_iso3
- location_code
- border
- age
- sex
- *_id
- *_tag

Usage

```
expand_time_to(
  x,
  max_isoyear = NULL,
  max_isoyearweek = NULL,
  max_date = NULL,
  ...
)
```

Arguments

x	An object of type csfmt_rts_data_v1
max_isoyear	Maximum isoyear
max_isoyearweek	Maximum isoyearweek
max_date	Maximum date
...	Not used.

Value

csfmt_rts_data_v1, a larger dataset that includes more rows corresponding to more time.

See Also

Other csfmt_rts_data: `identify_data_structure()`, `remove_class_csfmt_rts_data()`, `set_csfmt_rts_data_v1()`, `unique_time_series()`

`generate_test_data` *Generate test data*

Description

Generates some test data

Usage

```
generate_test_data(fmt = "csfmt_rts_data_v1")
```

Arguments

`fmt`	Data format (`csfmt_rts_data_v1`)

Value

csfmt_rts_data_v1, a dataset containing fake data.

Examples

```
cstidy::generate_test_data("csfmt_rts_data_v1")
```

`heal_time_csfmt_rts_data_v1` *Provides corresponding healed times*

Description

Provides corresponding healed times

Usage

```
heal_time_csfmt_rts_data_v1(x, cols, granularity_time = "date")
```

Arguments

x	A vector containing either dates, isoyearweek, or isoyear.
cols	Columns to restrict the output to.
granularity_time	date, isoyearweek, or isoyear, depending on the values contained in x.

Value

data.table, a dataset with time columns corresponding to the values given in x.

---

identify_data_structure

Hash the data structure of a dataset for a given column

Description

Reduces the data structure of a column inside a dataset into something that describes

Usage

```
identify_data_structure(x, col, ...)

## S3 method for class 'csfmt_rts_data_v1'
identify_data_structure(x, col, ...)

## S3 method for class '`tbl_Microsoft SQL Server`'
identify_data_structure(x, col, ...)
```

Arguments

x	An object
col	Column name to hash
...	Arguments passed to or from other methods

Value

csfmt_rts_data_structure_hash_v1, a summary object.

See Also

Other csfmt_rts_data: `expand_time_to()`, `remove_class_csfmt_rts_data()`, `set_csfmt_rts_data_v1()`, `unique_time_series()`

Examples

```
cstidy::generate_test_data() %>%
  cstidy::set_csfmt_rts_data_v1() %>%
  cstidy::identify_data_structure("deaths_n") %>%
  plot()
```

`nor_covid19_cases_by_time_location_csfmt_rts_v1`
Covid-19 data for PCR-confirmed cases in Norway (nation and county)

Description

This data comes from the Norwegian Surveillance System for Communicable Diseases (MSIS). The date corresponds to when the PCR-test was taken.

Usage

```
nor_covid19_cases_by_time_location_csfmt_rts_v1
```

Format

A csfmt_rts_data_v1 with 11028 rows and 18 variables:

granularity_time day/isoweek

granularity_geo nation, county

country_iso3 nor

location_code norge, 11 counties

border 2020

age total

isoyear Isoyear of event

isoweek Isoweek of event

isoyearweek Isoyearweek of event

season Season of event

seasonweek Seasonweek of event

calyear Calyear of event

calmonth Calmonth of event

calyearmonth Calyearmonth of event

date Date of event

covid19_cases_testdate_n Number of confirmed covid19 cases

covid19_cases_testdate_pr100000 Number of confirmed covid19 cases per 100.000 population

Details

The raw number of cases and cases per 100.000 population are recorded.

This data was extracted on 2022-05-04.

Source

https://github.com/folkehelseinstituttet/surveillance_data/blob/master/covid19/_DOCUMENTATION_data_covid19_msis_by_time_location.txt

---

nor_covid19_icu_and_hospitalization_csfmt_rts_v1

Norwegian Covid-19 data for ICU and hospitalization

Description

This data was extracted on 2022-05-04.

Usage

```
nor_covid19_icu_and_hospitalization_csfmt_rts_v1
```

Format

A csfmt_rts_data_v1 with 919 rows and 18 variables:

granularity_time day/isoweek

granularity_geo nation

country_iso3 nor

location_code norge

border 2020

age total

isoyear Isoyear of event

isoweek Isoweek of event

isoyearweek Isoyearweek of event

season Season of event

seasonweek Seasonweek of event

calyear Calyear of event

calmonth Calmonth of event

calyearmonth Calyearmonth of event

date Date of event

icu_with_positive_pcr_n Number of new admissions to the ICU with a positive PCR test

hospitalization_with_covid19_as_primary_cause_n Number of new hospitalizations with Covid-19 as the primary cause

Source

https://github.com/folkehelseinstituttet/surveillance_data/blob/master/covid19/_DOCUMENTATION_data_covid19_hospital_by_time.txt

---

`remove_class_csfmt_rts_data` *Remove class csfmt_rts_data_**

Description

Remove class csfmt_rts_data_*

Usage

```
remove_class_csfmt_rts_data(x)
```

Arguments

x	data.table

Value

No return value, called for the side effect of removing the csfmt_rts_data class from x.

See Also

Other csfmt_rts_data: `expand_time_to()`, `identify_data_structure()`, `set_csfmt_rts_data_v1()`, `unique_time_series()`

Examples

```
x <- cstidy::generate_test_data() %>%
  cstidy::set_csfmt_rts_data_v1()
class(x)
cstidy::remove_class_csfmt_rts_data(x)
class(x)
```

`set_csfmt_rts_data_v1` *Convert data.table to csfmt_rts_data_v1*

Description

`set_csfmt_rts_data_v1` converts a `data.table` to `csfmt_rts_data_v1` by reference. `csfmt_rts_data_v1` creates a new `csfmt_rts_data_v1` (not by reference) from either a `data.table` or `data.frame`.

Usage

```
set_csfmt_rts_data_v1(x, create_unified_columns = TRUE, heal = TRUE)

csfmt_rts_data_v1(x, create_unified_columns = TRUE, heal = TRUE)
```

Arguments

`x`	The data.table to be converted to csfmt_rts_data_v1
`create_unified_columns`	Do you want it to create unified columns?
`heal`	Do you want to impute missing values on creation?

Details

For more details see the vignette: `vignette("csfmt_rts_data_v1", package = "cstidy")`

Value

An extended `data.table`, which has been modified by reference and returned (invisibly).

No return value, called for side effect of replacing the current data.table with a csfmt_rts_data_v1 in place.

Returns a duplicated csfmt_rts_data_v1.

Smart assignment

`csfmt_rts_data_v1` contains the smart assignment feature for time and geography.

When the **variables in bold** are assigned using `:=`, the listed variables will be automatically imputed.

location_code:

- granularity_geo
- country_iso3

isoyear:

- granularity_time
- isoweek

- isoyearweek
- season
- seasonweek
- calyear
- calmonth
- calyearmonth
- date

isoyearweek:

- granularity_time
- isoyear
- isoweek
- season
- seasonweek
- calyear
- calmonth
- calyearmonth
- date

date:

- granularity_time
- isoyear
- isoweek
- isoyearweek
- season
- seasonweek
- calyear
- calmonth
- calyearmonth

Unified columns

`csfmt_rts_data_v1` contains 16 unified columns:

- granularity_time
- granularity_geo
- country_iso3
- location_code
- border
- age

- sex
- isoyear
- isoweek
- isoyearweek
- season
- seasonweek
- calyear
- calmonth
- calyearmonth
- date

See Also

Other csfmt_rts_data: `expand_time_to()`, `identify_data_structure()`, `remove_class_csfmt_rts_data()`, `unique_time_series()`

Examples

```
# Create some fake data as data.table
d <- cstidy::generate_test_data(fmt = "csfmt_rts_data_v1")
d <- d[1:5]

# convert to csfmt_rts_data_v1 by reference
cstidy::set_csfmt_rts_data_v1(d, create_unified_columns = TRUE)

#
d[1, isoyearweek := "2021-01"]
d
d[2, isoyear := 2019]
d
d[3, date := as.Date("2020-01-01")]
d
d[4, c("isoyear", "isoyearweek") := .(2021, "2021-01")]
d
d[5, c("location_code") := .("norge")]
d

# Investigating the data structure of one column inside a dataset
cstidy::generate_test_data() %>%
  cstidy::set_csfmt_rts_data_v1() %>%
  cstidy::identify_data_structure("deaths_n") %>%
  plot()
# Investigating the data structure via summary
cstidy::generate_test_data() %>%
  cstidy::set_csfmt_rts_data_v1() %>%
  summary()
```

`unique_time_series` *Unique time series*

Description

Attempts to identify the unique time series that exist in this dataset.

A time series is defined as a unique combination of:

- granularity_time
- granularity_geo
- country_iso3
- location_code
- border
- age
- sex
- *_id
- *_tag

Usage

```
unique_time_series(x, set_time_series_id = FALSE, ...)
```

Arguments

`x`	An object of type `csfmt_rts_data_v1`
`set_time_series_id`	If TRUE, then `x` will have a new column called 'time_series_id'
`...`	Not used.

Value

data.table, a dataset that lists all the unique time series in x.

See Also

Other csfmt_rts_data: `expand_time_to()`, `identify_data_structure()`, `remove_class_csfmt_rts_data()`, `set_csfmt_rts_data_v1()`

Index