Package 'readabs'

November 20, 2020

Type Package

Title Download and Tidy Time Series Data from the Australian Bureau of Statistics

Version 0.4.6

Maintainer Matt Cowgill <mattcowgill@gmail.com>

Description Downloads, imports, and tidies time series data from the Australian Bureau of Statistics <https://www.abs.gov.au/>.

Date 2020-11-20

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Depends R (>= 3.5)

Imports readxl (>= 1.2.0), tibble (>= 1.4.99), dplyr (>= 0.8.0), hutils (>= 1.5.0), fst, curl, purrr, tidyr (>= 1.0.0), stringr, stringi, rsdmx, tools, glue, httr, rvest, xml2, rlang

URL https://github.com/mattcowgill/readabs

BugReports https://github.com/mattcowgill/readabs/issues

RoxygenNote 7.1.1

VignetteBuilder knitr

Suggests knitr, rmarkdown, testthat (>= 2.1.0), RCurl, ggplot2

NeedsCompilation no

Author Matt Cowgill [aut, cre],
Zoe Meers [aut],
Jaron Lee [aut],
David Diviny [aut],
Hugh Parsonage [ctb]

Repository CRAN

Date/Publication 2020-11-20 10:50:05 UTC

R topics documented:

`download_abs_data_cube` *Experimental helper function to download ABS data cubes that are not compatible with read_abs.*

Description

`download_abs_data_cube()` downloads the latest ABS data cubes based on the catalogue name (from the new website url) and cube. The function downloads the file to disk. In comparison to `read_abs()` this function doesn't tidy the data.

Usage

```
download_abs_data_cube(
  catalogue_string,
  cube,
  path = Sys.getenv("R_READABS_PATH", unset = tempdir())
)
```

Arguments

`catalogue_string` ABS catalogue name as a string from the new website. For example, Labour Force, Australia, Detailed is "labour-force-australia-detailed". The possible catalogues can be obtained using the helper function `show_available_catalogues()`

cube	character. A character string that is either the complete filename or (uniquely) in the filename of the data cube you want to download, e.g. "EQ09". #' The available filenames can be obtained using the helper function `get_available_filenames()`
path	Local directory in which downloaded files should be stored. By default, 'path' takes the value set in the #' environment variable "R_READABS_PATH". If this variable is not set, #' any files downloaded by read_abs() will be stored in a temporary directory #' (`tempdir()`). See `Details` below for #' more information.

Details

'download_abs_data_cube()' downloads a file from the ABS containing a data cube. These files need to be saved somewhere on your disk. This local directory can be controlled using the 'path' argument to 'read_abs()'. If the 'path' argument is not set, 'read_abs()' will store the files in a directory set in the "R_READABS_PATH" environment variable. If this variable isn't set, files will be saved in a temporary directory.

To check the value of the "R_READABS_PATH" variable, run `Sys.getenv("R_READABS_PATH")`. You can set the value of this variable for a single session using `Sys.setenv(R_READABS_PATH = <path>)`. If you would like to change this variable for all future R sessions, edit your '.Renviron' file and add `R_READABS_PATH = <path>` line. The easiest way to edit this file is using `usethis::edit_r_environ()`.

The filepath is returned invisibly which enables piping to `unzip()` or `readxl::read_excel`.

Examples

```
## Not run:
download_abs_data_cube(
  catalogue_string = "labour-force-australia-detailed",
  cube = "EQ09"
)

## End(Not run)
```

---

extract_abs_sheets *Extract data sheets from an ABS timeseries workbook saved locally as an Excel file.*

Description

Note that this function will not tidy the data for you. Use 'read_abs_local()'to import and tidy data from local ABS time series spreadsheets or 'read_abs()' to download, import and tidy ABS time series.

Usage

```
extract_abs_sheets(
  filename,
  table_title = NULL,
  path = Sys.getenv("R_READABS_PATH", unset = tempdir())
)
```

Arguments

filename	Filename for an ABS time series spreadsheet (as string)
table_title	String giving the full title of the ABS table, such as "Table 1. Employed persons, Australia"
path	Local directory in which an ABS time series is stored. Default is 'Sys.getenv("R_READABS_PATH", unset = tempdir())'.

---

get_available_files *Helper function for* download_abs_data_cube *to show the available catalogues.*

Description

This function lists the possible files that are available in a catalogue. The filename (or an unambiguous part of the filename) must be specified as a string as an argument to download_abs_data_cube.

Usage

```
get_available_files(catalogue_string, refresh = FALSE)
```

Arguments

catalogue_string	character string specifying the catalogue, e.g. "labour-force-australia-detailed". You can use show_available_catalogues to find this out.
refresh	logical; 'FALSE' by default. If 'FALSE', an internal table of the available ABS catalogues is used. If 'TRUE', this table is refreshed from the ABS website.

Value

A tibble containing the title of the file, the filename and the complete url.

Examples

```
## Not run:
get_available_files("labour-force-australia-detailed")

## End(Not run)
```

read_abs *Download, extract, and tidy ABS time series spreadsheets*

Description

read_abs() downloads ABS time series spreadsheets, then extracts the data from those spreadsheets, then tidies the data. The result is a single data frame (tibble) containing tidied data.

Usage

```
read_abs(
  cat_no = NULL,
  tables = "all",
  series_id = NULL,
  path = Sys.getenv("R_READABS_PATH", unset = tempdir()),
  metadata = TRUE,
  show_progress_bars = TRUE,
  retain_files = TRUE,
  check_local = TRUE
)
```

Arguments

cat_no	ABS catalogue number, as a string, including the extension. For example, "6202.0".
tables	numeric. Time series tables in 'cat_no" to download and extract. Default is "all", which will read all time series in 'cat_no'. Specify 'tables' to download and import specific tables(s) - eg. 'tables = 1' or 'tables = c(1, 5)'.
series_id	(optional) character. Supply an ABS unique time series identifier (such as "A2325807L") to get only that series. This is an alternative to specifying 'cat_no'.
path	Local directory in which downloaded ABS time series spreadsheets should be stored. By default, 'path' takes the value set in the environment variable "R_READABS_PATH". If this variable is not set, any files downloaded by read_abs() will be stored in a temporary directory (tempdir()). See Details below for more information.
metadata	logical. If 'TRUE' (the default), a tidy data frame including ABS metadata (series name, table name, etc.) is included in the output. If 'FALSE', metadata is dropped.
show_progress_bars	TRUE by default. If set to FALSE, progress bars will not be shown when ABS spreadsheets are downloading.
retain_files	when TRUE (the default), the spreadsheets downloaded from the ABS website will be saved in the directory specified with 'path'. If set to 'FALSE', the files will be stored in a temporary directory.
check_local	If 'TRUE', the default, local 'fst' files are used, if present.

Details

'read_abs()' downloads spreadsheet(s) from the ABS containing time series data. These files need to be saved somewhere on your disk. This local directory can be controlled using the 'path' argument to 'read_abs()'. If the 'path' argument is not set, 'read_abs()' will store the files in a directory set in the "R_READABS_PATH" environment variable. If this variable isn't set, files will be saved in a temporary directory.

To check the value of the "R_READABS_PATH" variable, run `Sys.getenv("R_READABS_PATH")`. You can set the value of this variable for a single session using `Sys.setenv(R_READABS_PATH = <path>)`. If you would like to change this variable for all future R sessions, edit your '.Renviron' file and add `R_READABS_PATH = <path>` line. The easiest way to edit this file is using `usethis::edit_r_environ()`.

Value

A data frame (tibble) containing the tidied data from the ABS time series table(s).

Examples

```
# Download and tidy all time series spreadsheets
# from the Wage Price Index (6345.0)
## Not run:
wpi <- read_abs("6345.0")

## End(Not run)

# Get two specific time series, based on their time series IDs
## Not run:
cpi <- read_abs(series_id = c("A2325806K", "A2325807L"))

## End(Not run)
```

`read_abs_data` *Extracts ABS time series data from local Excel spreadsheets and converts to long format.*

Description

'read_abs_data()' is soft deprecated and will be removed in a future version. Please use 'read_abs_local()' to import and tidy locally-stored ABS time series spreadsheets, or 'read_abs()' to download, import, and tidy time series spreadsheets from the ABS website.

Usage

```
read_abs_data(path, sheet)
```

Arguments

path	Filepath to Excel spreadsheet.
sheet	Sheet name or number.

Value

Long-format dataframe

---

read_abs_local *Read and tidy locally-saved ABS time series spreadsheet(s)*

Description

If you need to download and tidy time series data from the ABS, use `read_abs()`. `read_abs_local()` imports and tidies data from ABS time series spreadsheets that are already saved to your local drive.

Usage

```
read_abs_local(
  cat_no = NULL,
  filenames = NULL,
  path = Sys.getenv("R_READABS_PATH", unset = tempdir()),
  use_fst = TRUE,
  metadata = TRUE
)
```

Arguments

cat_no	character; a single catalogue number such as "6202.0". When ‘cat_no‘ is specified, all local files in ‘path‘ corresponding to the specified catalogue number will be imported. For example, if you run ‘read_abs_local("6202.0")‘, it will look in "data/ABS/6202.0" and attempt to load any .xls files in that location. If ‘cat_no‘‘ is specified, ‘filenames‘ will be ignored.
filenames	character vector of at least one filename of a locally-stored ABS time series spreadsheet. For example, "6202001.xls" or c("6202001.xls", "6202005.xls"). Ignored if a value is supplied to ‘cat_no‘. If ‘filenames‘ is blank and ‘cat_no‘ is blank, ‘read_abs_local()‘ will attempt to read all .xls files in the directory specified with ‘path‘.
path	path to local directory containing ABS time series file(s). Default is ‘Sys.getenv("R_READABS_PATH", unset = tempdir())‘. If nothing is specified in ‘filenames‘ or ‘cat_no‘, ‘read_abs_local()‘ will attempt to read all .xls files in the directory specified with ‘path‘.
use_fst	logical. If ‘TRUE‘ (the default) then, if an ‘fst‘ file of the tidy data frame has already been saved in ‘path‘, it is read immediately.
metadata	logical. If ‘TRUE‘ (the default), a tidy data frame including ABS metadata (series name, table name, etc.) is included in the output. If ‘FALSE‘, metadata is dropped.

Details

Unlike `read_abs()`, the 'table_title' column in the data frame returned by `read_abs_local()` is blank. If you require 'table_title', please use `read_abs()` instead.

Examples

```
# Load and tidy two specified files from the "data/ABS" subdirectory
# of your working directory
## Not run:
lfs <- read_abs_local(c("6202001.xls", "6202005.xls"))

## End(Not run)
```

`read_abs_metadata` *Extracts ABS series metadata directly from Excel spreadsheets and converts to long-form.*

Description

Extracts ABS series metadata directly from Excel spreadsheets and converts to long-form.

Usage

```
read_abs_metadata(path, sheet)
```

Arguments

`path`	Filepath to Excel spreadsheet.
`sheet`	Sheet name or number.

Value

Long-form dataframe

read_abs_sdmx *Extracts ABS XML-formatted data using the SDMX API*

Description

Access the sdmx URLs at 'http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1407.0.55.002Main+Features4User+Guide'

Usage

```
read_abs_sdmx(url)
```

Arguments

url	URL weblink.

Value

data frame

Examples

```
## Not run:
url <- paste0(
  "http://stat.data.abs.gov.au/restsdmx/sdmx.ashx/GetData/LF/",
  "0.2+3+4+11+13+6+15+14+10.3+1+2.1519+1599.10+20+30.M/",
  "all?startTime=2017-12&endTime=2018-11"
)
lfs <- read_abs_sdmx(url)
lfs

## End(Not run)
```

read_awe *read_awe*

Description

Convenience function to obtain wage levels from ABS 6302.0, Average Weekly Earnings, Australia.

Usage

```
read_awe(
  wage_measure = c("awote", "ftawe", "awe"),
  sex = c("persons", "males", "females"),
  na.rm = FALSE,
  path = Sys.getenv("R_READABS_PATH", unset = tempdir()),
  show_progress_bars = FALSE,
  check_local = FALSE
)
```

Arguments

wage_measure	Character. Must be one of: • 'awote' Average weekly ordinary time earnings; also known as Full-time adult ordinary time earnings • 'ftawe' Full-time adult total earnings • 'awe' Average weekly total earnings of all employees
sex	Character. Must be one of: 'persons', 'males', or 'females'.
na.rm	Logical. 'FALSE' by default. If 'FALSE', a consistent quarterly series is returned, with 'NA' values for quarters in which there is no data. If 'TRUE', only dates with data are included in the returned data frame.
path	See '?read_abs'
show_progress_bars	See '?read_abs'
check_local	See '?read_abs'

Details

The latest AWE data is available using 'read_abs(cat_no = "6302.0", tables = 2)'. However, this time series only goes back to 2012, when the ABS switched from quarterly to biannual collection and release of the AWE data. The 'read_awe()' function assembles on time series back to November 1983 quarter; it is quarterly to 2012 and biannual from then. Note that the data returned with this function is consistently quarterly; any quarters for which there are no observations are recorded as 'NA' unless 'na.rm' = 'TRUE'.

Value

A 'tbl_df' with four columns: 'date', 'sex', 'wage_measure' and 'value'. The data is nominal (ie. not inflation-adjusted).

Examples

```
## Not run:
read_awe("awote", "persons")

## End(Not run)
```

read_cpi *Download a tidy tibble containing the Consumer Price Index from the ABS*

Description

read_cpi() uses the read_abs() function to download, import, and tidy the Consumer Price Index from the ABS. It returns a tibble containing two columns: the date and the CPI index value that corresponds to that date. This makes joining the CPI to another dataframe easy. read_cpi() returns the original (ie. not seasonally adjusted) all groups CPI for Australia. If you want the analytical series (eg. seasonally adjusted CPI, or trimmed mean CPI), you can use read_abs().

Usage

```
read_cpi(
  path = Sys.getenv("R_READABS_PATH", unset = tempdir()),
  show_progress_bars = TRUE,
  check_local = FALSE,
  retain_files = FALSE
)
```

Arguments

path	character; default is "data/ABS". Only used if retain_files is set to TRUE. Local directory in which to save downloaded ABS time series spreadsheets.
show_progress_bars	logical; TRUE by default. If set to FALSE, progress bars will not be shown when ABS spreadsheets are downloading.
check_local	logical; FALSE by default. See ?read_abs.
retain_files	logical; FALSE by default. When TRUE, the spreadsheets downloaded from the ABS website will be saved in the directory specified with 'path'.

Examples

```
# Create a tibble called 'cpi' that contains the CPI index
# numbers for each quarter

cpi <- read_cpi()


# This tibble can now be joined to another to help streamline the process of
# deflating nominal values.
```

`read_payrolls` *Download and tidy ABS payroll jobs and wages data*

Description

Import a tidy tibble of ABS Weekly Payrolls data.

Usage

```
read_payrolls(
  series = c("industry_jobs", "industry_wages", "sa4_jobs", "sa3_jobs",
    "subindustry_jobs", "empsize_jobs"),
  path = Sys.getenv("R_READABS_PATH", unset = tempdir())
)
```

Arguments

`series`	Character. Must be one of: • "industry_jobs" Payroll jobs by industry division, state, sex, and age group (Table 4) • "industry_wages" Total wages by industry division, state, sex, and age group (Table 4) • "sa4_jobs" Payroll jobs by statistical area 4 (SA4) and state (Table 5) • "sa3_jobs Payroll jobs by statistical area 4 (SA4), statistical area 3 (SA3), and state (Table 5) • "subindustry_jobs" Payroll jobs by industry sub-division and industry division (Table 6) • "empsize_jobs" Payroll jobs by size of employer (number of employees) and state (Table 7) The default is "industry_jobs".
`path`	Local directory in which downloaded ABS time series spreadsheets should be stored. By default, 'path' takes the value set in the environment variable "R_READABS_PATH". If this variable is not set, any files downloaded by read_abs() will be stored in a temporary directory (`tempdir()`).

Details

The ABS 'Weekly Payroll Jobs and Wages in Australia' dataset is very useful to analysts of the Australian labour market. It draws upon data collected by the Australian Taxation Office as part of its Single-Touch Payroll initiative and supplements the monthly Labour Force Survey. Unfortunately, the data as published by the ABS (1) is not in a standard time series spreadsheet; and (2) is messy in various ways that make it hard to read in R. This convenience function uses 'download_abs_data_cube()' to import the payrolls data, and then tidies it up.

Value

A tidy (long) 'tbl_df'. The number of columns differs based on the 'series'.

Examples

```
## Not run:
# Fetch payroll jobs by industry and state (the default, "industry_jobs")
read_payrolls()

# Payroll jobs by employer size
read_payrolls("empsize_jobs")

## End(Not run)
```

`scrape_abs_catalogues` *Helper function for* `download_abs_data_cube` *to scrape the available catalogues from the ABS website.*

Description

This function downloads a new version of the lookup table used by `show_available_catalogues`.

Usage

```
scrape_abs_catalogues()
```

Value

A tibble containing the catalogues and how they are organised on the ABS website.

`separate_series` *Separate the series column in a tidy ABS time series data frame*

Description

Separate the 'series' column in a data frame (tibble) downloaded using `read_abs()` into multiple columns using the ";" separator.

Usage

```
separate_series(
  data,
  column_names = NULL,
  remove_totals = FALSE,
  remove_nas = FALSE
)
```

Arguments

data	A data frame (tibble) containing tidied data from the ABS time series table(s).
column_names	(optional) character vector. Supply a vector of column names, such as c("group_name","variable","ge If not supplied, columns will be named "series_1" etc.
remove_totals	logical. FALSE by default. If set to TRUE, any series rows that contain the word "total" will be removed.
remove_nas	logical. FALSE by default. If set to TRUE, any rows containining an NA in at least one of the separated series columns will be removed.

Value

A data frame (tibble) containing the tidied data from the ABS time series table(s).

Examples

```
## Not run:
motor_vehicles <- read_abs("9314.0") %>%
  separate_series()

## End(Not run)
```

---

show_available_catalogues

Helper function for download_abs_data_cube *to show the available catalogues.*

Description

This function lists the possible catalogues that are available on the ABS website. These catalogues must be specified as a string as an argument to download_abs_data_cube.

Usage

```
show_available_catalogues(selected_heading = NULL, refresh = FALSE)
```

Arguments

selected_heading	optional character string specifying the heading on the ABS statistics webpage. e.g. "Earnings and work hours"
refresh	logical; 'FALSE' by default. If 'FALSE', an internal table of the available ABS catalogues is used. If 'TRUE', this table is refreshed from the ABS website.

Value

a character vector of catalogues.

Examples

```
show_available_catalogues("Earnings and work hours")
```

`tidy_abs` *Tidy ABS time series data.*

Description

Tidy ABS time series data.

Usage

```
tidy_abs(df, metadata = TRUE)
```

Arguments

`df`	A data frame containing ABS time series data that has been extracted using `extract_abs_sheets`.
`metadata`	logical. If 'TRUE' (the default), a tidy data frame including ABS metadata (series name, table name, etc.) is included in the output. If 'FALSE', metadata is dropped.

Value

data frame (tibble) in long format.

Examples

```
# First extract the data from the local spreadsheet
## Not run:
wpi <- extract_abs_sheets("634501.xls")

## End(Not run)

# Then tidy the data extracted from the spreadsheet. Note that
# \code{extract_abs_sheets()} returns a list of data frames, so we need to
# subset the list.
## Not run:
tidy_wpi <- tidy_abs(wpi[[1]])

## End(Not run)
```

`tidy_abs_list` *Tidy multiple dataframes of ABS time series data contained in a list.*

Description

Tidy multiple dataframes of ABS time series data contained in a list.

Usage

```
tidy_abs_list(list_of_dfs, metadata = TRUE)
```

Arguments

`list_of_dfs`	A list of dataframes containing extracted ABS time series data.
`metadata`	logical. If 'TRUE' (the default), a tidy data frame including ABS metadata (series name, table name, etc.) is included in the output. If 'FALSE', metadata is dropped.

Index