Package 'datazoom.social'

July 6, 2026

Title Simplify Access to Brazilian Social Data

Version 0.1.1

Description Provides tools for downloading and processing microdata from the PNAD Contínua (PNADC, Continuous National Household Sample Survey), a rotating panel survey published quarterly by IBGE (Brazilian Institute of Geography and Statistics). Includes panel identification algorithms for linking individuals across survey waves.

License MIT + file LICENSE

URL https://datazoom.com.br/en/

Imports arrow, dplyr, igraph, magrittr, PNADcIBGE, purrr, rlang, stringr, tidyr

Encoding UTF-8

Suggests knitr, rmarkdown

VignetteBuilder knitr

Depends R (>= 4.1.0)

BugReports https://github.com/datazoompuc/datazoom.social/issues

LazyData true

Config/roxygen2/version 8.0.0

NeedsCompilation no

Author Laura Tavares Regadas [aut, cre], DataZoom (PUC-Rio) [fnd], Igor Rigolon Veiga [aut], Arthur Lins de Vasconcellos [aut], Giulia Toscano Imbuzeiro [aut], Guilherme Jardim [aut], Pablo Chaves [aut], Breno Avidos [aut], Bernardo Sieira [aut]

Maintainer Laura Tavares Regadas <lauratregadas@gmail.com>

Repository CRAN

Date/Publication 2026-07-06 15:40:02 UTC

Contents

`build_pnadc_panel` *Build PNADc Panel*

Description

This function builds a panel dataset from PNADC data, identifying households and individuals.

Usage

```
build_pnadc_panel(dat, panel)
```

Arguments

`dat`	Data frame with PNADC data, sorted into a single panel.
`panel`	A `character` with the type of panel identification. Use "none" for no paneling, "basic" for basic paneling, "advanced_1" for advanced stage 1 paneling, "advanced_2" for advanced stage 2 paneling, and "advanced_3" for the fuzzy-matching stage 3 paneling.

Value

A modified dataset with added identifiers for household (`id_dom`) and individual (`id_ind`, and progressively `id_rs1`, `id_rs2`, or `id_rs3`) based on the chosen panel algorithm.

Examples

```
# Example usage:

panel_data <- build_pnadc_panel(dat = pnad_sample, panel = "advanced_3")
```

cria_df_de_atrito *Create an attrition table for a panel file*

Description

This function generates a summary dataframe indicating the count of missing interviews for each individual and the unconditional tracking rates.

Usage

```
cria_df_de_atrito(data, panel)
```

Arguments

data	The input data frame, preferably a PNADc panel file.
panel	The identification strategy: "basic", "advanced_1", "advanced_2", "advanced_3" or "households".

Value

A data frame summarizing missing interviews and the tracking rates.

load_pnadc *Load Continuous PNAD Data*

Description

This function downloads PNADC data and applies panel identification algorithms

Usage

```
load_pnadc(
  save_to,
  years,
  quarters = 1:4,
  panel = "advanced",
  raw_data = FALSE,
  save_options = c(TRUE, TRUE),
  vars = NULL
)
```

Arguments

save_to	A character with the directory in which to save the downloaded files.
years	A numeric indicating for which years the data will be loaded, in the format YYYY. Can be any vector of numbers, such as 2010:2012.
quarters	The quarters within those years to be downloaded. Can be a numeric vector or a list of vectors, for different quarters per year.
panel	A character choosing the panel algorithm to apply ("none", "basic", or "advanced"). For details, check vignette("BUILD_PNADC_PANEL")
raw_data	A logical setting the return of raw (TRUE) or processed (FALSE) variables.
save_options	A logical vector of length 2. Controls whether quarterly files are saved and in which format all files are saved. Panel files are always saved. There are four possible combinations: • c(TRUE, TRUE): saves quarterly and panel files in .rds format. This is the default. • c(TRUE, FALSE): saves quarterly and panel files in .parquet format. • c(FALSE, TRUE): does not save quarterly files; panel files are saved in .rds format. • c(FALSE, FALSE): does not save quarterly files; panel files are saved in .parquet format.
vars	A character vector of additional variable names to be downloaded, following the same convention as the vars parameter in get_pnadc. Each name must match a column in the PNADC microdata exactly as published by IBGE (e.g. "VD4019", "V2009"). Note that get_pnadc always returns a set of structural columns regardless of this argument, these include survey design weights (V1027, V1028, V1028001, V1028200, posest, posest_sxi), deflator variables (Habitual, Efetivo), and identifiers such as UF, Estrato, V1029, V1033, ID_DOMICILIO, totalling around 233 columns. The vars argument adds *on top of* those columns; it does not restrict them. Use NULL (the default) to download all available microdata columns. If panel is not "none", any columns required by the panel identification algorithm that are missing from vars will be added automatically and a warning will list the columns that were added. The required columns per algorithm are: • "basic": UPA, V1008, V1014, V2007, V20082, V20081, V2008. • "advanced": all of the above, plus V2003. Note that several of these (UPA, V1008, V1014) are part of the structural columns always returned by get_pnadc, so in practice only V2007, V20082, V20081, V2008 (and V2003 for "advanced") are likely to be auto-added.

Value

A message indicating the successful save of panel files.

Examples

```
### DO NOT RUN ###
load_pnadc(
  save_to = tempdir(),
  years = 2016,
  quarters = 1:4,
  panel = "advanced",
  raw_data = FALSE,
  save_options = c(FALSE, FALSE)
)
```

---

pnad_sample *Simulated PNAD sample dataset*

Description

A small simulated dataset inspired by microdata from the Brazilian Continuous National Household Sample Survey (PNAD Contínua), included for examples, tests, and documentation in the `datazoom.social` package.

Usage

```
pnad_sample
```

Format

A `data.table` and `data.frame` with 31 rows and 23 variables:

V1 Record identifier.

Ano Survey year.

Trimestre Survey quarter.

UF Federative unit code.

UPA Primary sampling unit identifier.

V1008 Household serial identifier.

V1014 Number of household members.

V1016 Household interview status or type code.

V20082 Year of birth.

V20081 Month of birth.

V2008 Day of birth or age-related auxiliary code, as provided in the simulated data.

V2007 Sex code.

V2009 Age in years.

VD3004 Educational attainment code.

VD4001 Labor force status code.

VD4002 Employment status code.

VD4005 Employment position or job category code.

VD4009 Usual hours worked category or related labor variable code.

VD4019 Monthly labor income.

V4010 Main job identifier or occupation-related code.

V4012 Economic activity or occupation grouping code.

V4013 Time in job or age-related auxiliary labor code.

V4022 Household or person weight, as represented in the simulated data.

Details

This dataset does not contain real PNAD observations. It was created only for demonstration purposes and includes a small subset of variables from the original files distributed by IBGE. Its reduced size makes package examples faster and lighter, while preserving a structure similar to that of the original survey data.

The purpose of `pnad_sample` is to provide a lightweight object that mimics part of the structure of PNAD Contínua microdata, allowing users to run examples without downloading or processing the full original files.

Variable names follow the naming convention used in the original survey microdata, but the values in this dataset are simulated and should not be used for substantive empirical analysis.

Source

Inspired by the structure of PNAD Contínua microdata produced by the Brazilian Institute of Geography and Statistics (IBGE). https://www.ibge.gov.br

Index