ID529: A short primer on data dictionaries

20th January 2023

Team

Sarah Baum

Heather Kelahan

Ruby Hickman

Sejeong Park

Diego Liang

Hodu Tesla

Lawson Ung

Agenda

- Why data dictionaries?
- Packages in R
- Best practices
- Try it yourself

Introduction

- Imagine you have inherited a dataset from a colleague named Hodu.

- Your mentor has asked you to conduct an analysis on this dataset.

- You are relieved because you know Hodu is a treasured friend and would never leave you in the lurch.
  - That is, he has left you a data dictionary.

- What does this data dictionary would include?

Hodu’s data dictionary contains…

- **Nature of the dataset**
  - Where did the dataset come from?
  - Who collected it?
  - Why was it collected?
  - Structure and format
  - Variables includes
  - Missing data
  - Key references, papers that have used these data, and further reading

- **Nature of the variables**
  - Type: are they binary, categorical, ordinal, string, dates, numeric, or other?
  - Range, allowed values, and units
  - Information on variables generated by the research team (e.g., if they have been categorized or transformed), with references if appropriate.
  - How missing data has been coded

Hodu’s data dictionary contains…

- **Nature of the dataset**
  - Where did the dataset come from?
  - Who collected it?
  - Why was it collected?
  - Structure and format
  - Variables includes
  - Missing data
  - Key references, papers that have used these data, and further reading

- **Nature of the variables**
  - Type: are they binary, categorical, ordinal, string, dates, numeric, or other?
  - Range, allowed values, and units
  - Information on variables generated by the research team (e.g., if they have been categorized or transformed), with references if appropriate.
  - How missing data has been coded

Hodu’s data dictionary contains…

The key is that anyone using the data and associated code should be able to:

- Reproduce your findings without having to make any assumptions about how variables have been treated, or why you reached certain analytic decisions

- Be able to use the data to conduct further analyses, if appropriate.

DETAIL, DETAIL, DETAIL

- Key references, papers that have used these data, and further reading

Packages in R

- *labelled*
- *codebook*
- *dataMeta*

labelled

Allows for the application and retrieval of variable labels, adding value labels, treating missing values, and generation of data dictionaries.

Key functions

- var_label(): apply or retrieve label(s)
  - Ex: var_label(data$var) <- “Variable label” will apply the label
  - Calling var_label(data$var) again will retrieve label
  - For tibbles/in a pipe: set_var_labs()
- labelled(): create a vector with labelled values
  - labelled(value_vector, value_labels_vector)
  - But note these vectors won’t work for analysis; must convert to factor, numeric, etc. before using them
  - For tibbles/in a pipe: set_value_labs()
- look_for(): generate a data dictionary
  - Variable, label, column type, and values are default outputs
  - Option details = TRUE generates more details

codebook

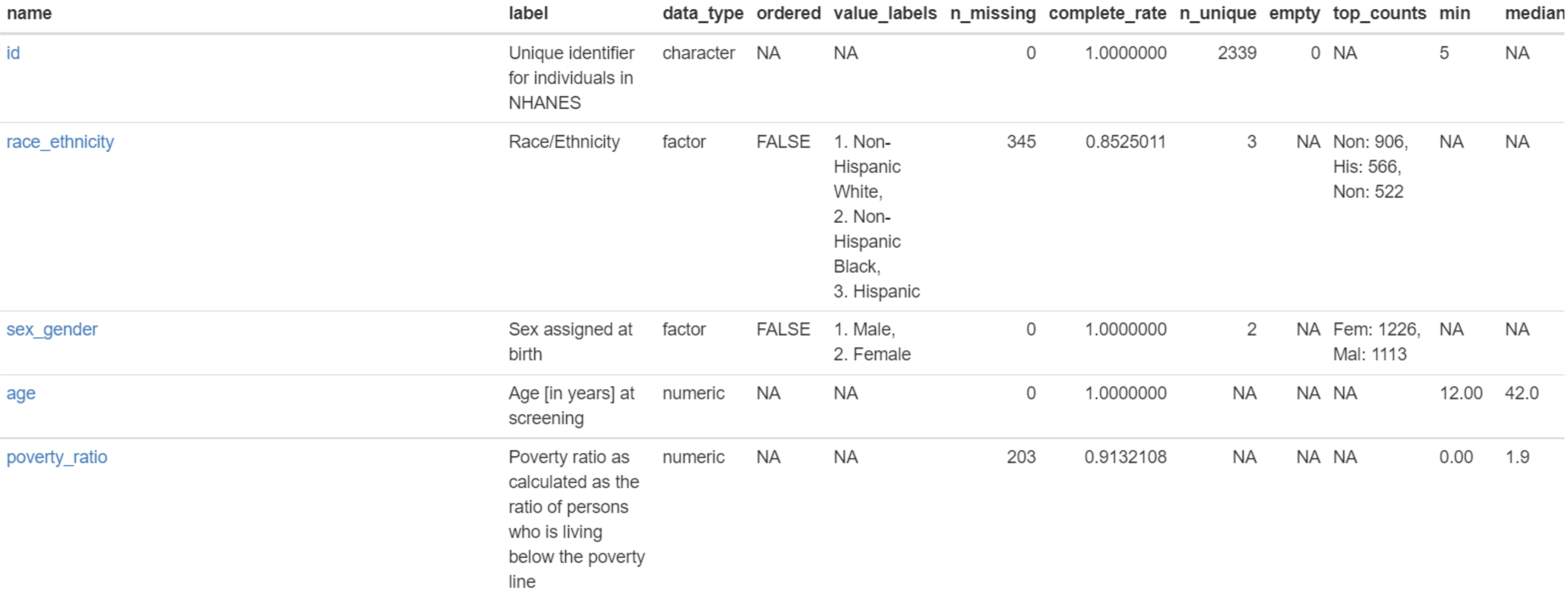
Automatically generates markdown codebook from your data frame

- Works well with the *labelled package* which you can use to manage variables
- Summarizes metadata, descriptive statistics, missing variables
- Allows you to modify labels and metadata
- Works with metadata from Stata and SPSS

Key functions

- *codebook::new_codebook_rmd()* - Write in console to launch new .rmd with defaults to generate codebook
- *codebook(your_data)* - generates full codebook
- *var_label()* - Uses functions from *labelled* package to modify attributes and labels

Codebook table



name	label	data_type	ordered	value_labels	n_missing	complete_rate	n_unique	empty	top_counts	min	median
id	Unique identifier for individuals in NHANES	character	NA	NA	0	1.0000000	2339	0	NA	5	NA
race_ethnicity	Race/Ethnicity	factor	FALSE	1. Non-Hispanic White, 2. Non-Hispanic Black, 3. Hispanic	345	0.8525011	3	NA	Non: 906, His: 566, Non: 522	NA	NA
sex_gender	Sex assigned at birth	factor	FALSE	1. Male, 2. Female	0	1.0000000	2	NA	Fem: 1226, Mal: 1113	NA	NA
age	Age [in years] at screening	numeric	NA	NA	0	1.0000000	NA	NA	NA	12.00	42.0
poverty_ratio	Poverty ratio as calculated as the ratio of persons who is living below the poverty line	numeric	NA	NA	203	0.9132108	NA	NA	NA	0.00	1.9

dataMeta

Functionalities

- **Linker**: an intermediary, contain the names of the variables, a description of each variable provided by the user and a "variable type."

- build_linker(): R will require that the user create two vectors that will fill out the variable descriptions and variable types.
- prompt_linker(): R will prompt the user to add the description of each variable in the console and the variable type.

Dictionary build: using build_dict()

dict <- build_dict (my.data = df, linker = linker, option_description = option_desc, prompt_varopts = FALSE)

- option_description: NULL or a vector object
- prompt_varopts: if "option_description" is not NULL, it must be FALSE; otherwise, R will prompt the user to add option description

Pearls of wisdom

- Data dictionaries are living documents, please update as you go (including on GitHub)
- More detail is better than not enough
- Depending on the size of the project, a flow chart can help document how data is flowing into the analysis, who is sharing it, and how often

Pearls of wisdom

- Document the data structure and relationship between your files
- Note any changes to the data over time, especially if reporting or coding of certain variables has changed.
- Be clear with your words and refer to existing dictionaries or common data elements in the topic area if needed.

Try it on your own

Please download our RMD tutorial from our GitHub repository, and try and exploring some of these packages and their functionalities.