|  |  |
| --- | --- |
| **CSDE 502 2021 Winter**  **Assignment 6**  ***Add Health data; Variable creation***  **Instructor: Phil Hurvitz**  [**phurvitz@uw.edu**](mailto:phurvitz@uw.edu) | **My Name:**  **My UWNetID:** |

**Due Date: 2021-02-18 09:00 AM**

**Instructions:**

1. Fill in your name and UWNetID above.
2. Put answers to the questions on this document, using the “00Answers” Word style so your answers are clearly distinguished from the questions.
3. Create a PDF file from this document.
4. Create a **single zip** file including this document as a PDF file, along with the RDS file and R code file.
5. Upload the **single zip file** to Canvas.

**Explanation:**

For this assignment, you will be perusing some of the documentation for the Add Health Wave 1 data set. You will use the documentation to make some updates to a data frame containing some of the Add Health data, and then save the data frame as an RDS file. You will update a metadata table that partially describes the data set and changes you made to the variable names and variable labels.

To open a Stata version 13 file in R there are two main options:

1. Use haven::read\_dta(). To access variable labels in R use labelled::foreign\_to\_labelled(). To update variable labels, use the labelled::var\_label() function.
2. Use readstata13::read.dta13(). Variable labels for this format are available, e.g., for a data frame named dat as attributes(dat)$var.labels. This is a vector of text strings that can be updated by assigning a new value to the specified element, e.g.,   
   attributes(dat)$var.labels[1] <- “foo”.

To save the RDS file, use the base function saveRDS().

Here is a base R code snippet that will rename a single variable:

colnames(data\_frame)[grep("^original\_variable\_name$", colnames(data\_frame))] <- new\_variable\_name

The grep() function finds the position of the named variable in the list of variables in the data frame. The characters ^ and $ are regular expressions to specify the start and end of the string to be matched (assuring that the pattern does not match multiple similar variable names).

It is much simpler with tidyverse and magrittr:

data\_frame %<>% rename(new\_variable\_name = old\_variable\_name)

Additional hint for dealing with PDF documentation:

1. Use pdfgrep (should be available in a Linux or Mac package manager; for Windows, search for a version or use Cygwin).
2. Use the R pdftools package. This could be used in a loop over each PDF file to create a data frame with the name of the PDF file, page number, and text of each page. The str\_match() function could be used to identify the file name and page number where specific text strings occur. For a minimal example, this shows that the string “h1gi1m” is found on page 1 of INH01PUB.PDF. Conversion of the PDF file’s text to lowercase simplifies the matching:

> x <- pdftools::pdf\_text(pdf = "INH01PUB.PDF")

> str\_match(string = x %>% str\_to\_lower(), pattern = "h1gi1m")

[,1]

[1,] "h1gi1m"

[2,] NA

[3,] NA

[4,] NA

[5,] NA

[6,] NA

[7,] NA

[8,] NA

[9,] NA

[10,] NA

[11,] NA

[12,] NA

[13,] NA

[14,] NA

[15,] NA

**Questions:**

1. Explore the Add Health website (<http://www.cpc.unc.edu/projects/addhealth>) and answer the following questions (making sure to cite as necessary):
   1. What was the sampling frame for this study?

The sampling frame for the Add Health study was all high schools included in the Quality Education Database (QED). High school was defined as schools with an 11th grade and more than 30 students.

* 1. What were the three kinds of respondents at Wave I?
  2. What was the instrument with the largest sample size?
  3. Is it possible for a respondent to be in Wave III without being in Wave II?
  4. What is the time span of the Add Health data collection (all waves)?
  5. What is the difference between the public and the restricted-use Add Health data?
  6. Describe a research question that you might be able to answer using the Add Health dataset.

1. Download the public-use Add Health documentation at <https://canvas.uw.edu/courses/1434040/files>. Answer the following questions:
   1. In what pdf document is the documentation for the race items for the Wave I In-Home questionnaire?
   2. How many respondents were of Hispanic/Latino origin?
   3. What is the "Knowledge Quiz" in the Wave I In-Home questionnaire?
   4. What is the unique identifier for the In-home data?
2. Download the Stata 13 format file AHwave1\_v1.dta (<http://staff.washington.edu/phurvitz/csde502_winter_2021/data/AHwave1_v1.dta>).
   1. Fill in the grey missing cells in Table 1 below based on the data and/or documentation. Optimally, use the documentation to familiarize yourself with the structure of the code books.
   2. Using questions 6 and 8 in INH01PUB.PDF, create a new variable named "race" that uses recoded values (white = 1; black/African American = 2; American Indian = 3; Asian/Pacific Islander = 4; other = 5; unknown/missing = 9).
   3. Rename the variables, and update variable labels using Table 1 as a guide and save the data frame as the file as **AHwave1\_v2.rds**. Use a single R code file for your edits to the data file.
   4. Update the status in Table 1 as needed.

Table 1: Codebook for variables from Add Health Wave 1 data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **new variable name** | **original variable name** | **status\*** | **data type** | **values** | **new  variable  label** | **codebook file name** |
| aid | aid | unchanged | text | 8 digit string | unique case (student) identifier | SECTAPUB.PDF |
| imonth | imonth | unchanged | integer | 1 4 to 12 | month interview completed | SECTAPUB.PDF |
| iday | iday | unchanged |  |  | day interview completed | SECTAPUB.PDF |
| iyear | iyear | unchanged |  | 94, 95 |  | SECTAPUB.PDF |
| bio\_sex | bio\_sex |  |  |  | interviewer confirmed sex |  |
| bmonth | h1gi1m |  |  |  | birth month | INH01PUB.PDF |
| byear | h1gi1y |  |  |  | birth year |  |
| hispanic | h1gi4 | renamed |  |  | Hispanic/Latino | INH01PUB.PDF |
| white | h1gi6a | renamed |  | 0 = not marked  1 = marked  6 = refused  8 = don't know | race white | INH01PUB.PDF |
| black |  |  |  |  | race black or African American | INH01PUB.PDF |
| AI | h1gi6c |  |  |  | race American Indian or Native American | INH01PUB.PDF |
| asian | h1gi6d |  |  |  | race Asian or Pacific Islander | INH01PUB.PDF |
| raceother | h1gi6e |  |  |  | race other | INH01PUB.PDF |
| onerace |  |  |  |  | one category best describes racial background | INH01PUB.PDF |
| observedrace | h1gi9 |  |  |  | interviewer observed race | INH01PUB.PDF |
| health | h1gh1 |  |  |  | how is your health |  |
| race | not applicable |  |  |  | race recoded as white; black/African American; American Indian; Asian/Pacific Islander; other; unknown/missing |  |

\*status categories: unchanged, renamed, missing defined, derived